UNIVERSITY OF CAMBRIDGE

Translational regulation in aggressive B-cell lymphomas

Joanna Alicja Krupka

Gonville and Caius College

This dissertation is submitted on Easter Term, 2021
for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the prescribed limit of 60 000 words.

Joanna Alicja Krupka
September, 2021

Abstract

Translational regulation in aggressive B-cell lymphomas

Joanna Alicja Krupka

The Germinal Centre (GC) reaction is a dynamic process where B-cells undergo recombination and somatic hypermutation of immunoglobulin genes in response to antigen stimulation. This essential component of the adaptive immune system is associated with cycles of intensive proliferation and selection, which carries a risk of malignant transformation. Aggressive lymphomas arising from the GC stage of B-cell development are the most common haematological malignancies with heterogeneous molecular mechanisms and clinical presentation. Although the last decade witnessed considerable advances in the biology of GC reaction and related tumours, the studies focused predominantly on the network of transcription factors.

The advances in Next Generation Sequencing technologies have opened new possibilities to explore mechanisms of regulation beyond the level of transcription. Ribo-Seq is a technique combining ribosome footprinting with deep sequencing of mRNA fragments that allows to map the position of translating ribosomes with single nucleotide precision.

Here I investigate the mechanisms of translational regulation contributing to lymphoma development. Firstly, I introduce RiboStream, an automated bioinformatic pipeline designed to streamline processing of Ribo-Seq datasets while maintaining transparency and reproducibility of the computational workflow. Then, I provide an overview and benchmarking of current methods for identifying translationally regulated genes. Based on these I select a strategy to reveal that overexpression of two B-cell oncogenes, BCL6 or MYC, is followed by preferential translational of selected transcripts. Next, I show that loss-of-function mutations in RNA-helicase (DDX3X) promote early development of MYC-driven lymphoma by buffering the effects of MYC on translation of ribosomal proteins and the rate of global protein synthesis. Finally, I explore a genome-wide distribution of translating ribosomes to study the scope of non-canonical translation in lymphoid cells. Taking advantage of a large dataset of 79 Ribo-Seq libraries I reveal pervasive translation of ostensibly non-coding regions, and design a knock-down CRISPR screen library to identify those important for B-cell survival.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor, Dr. Daniel Hodson, who took the risk of having me in his lab and gave me the freedom to discuss problems I found interesting. His expertise, advice and sense of humour were invaluable. I would like to thank Dan for being a trusted mentor and friend on my way to becoming a scientist.

I am also thankful to my second supervisor, Dr. Shamith Samarajiwa, for his enthusiasm for accompanying me with my first steps in bioinformatics. By sharing all computational resources with me, I enjoyed an unrestrained opportunity to explore my ideas, for which I am immensely grateful. I am also indebted to Dr. Martin Turner for inspiring discussions and encouragement to look ahead into unexplored territories of science.

I would like to also thank all members of the Hodson and the Samarajiwa labs, who made the last four years a good-humoured and colourful time. Special thanks go to Jie Gao for her strength of spirit in preparing all the Ribo-Seq libraries and to Chun Gong, Mata Vorri and Hendrik Runge for fruitful collaboration. My time in Cambridge would not be the same without Dora Bihary, Shoko Hirosue, Cassandra Kosmidou, David Shorthouse and Katie Young. Thank you for sharing your time with me and an entire series of pub events, which helped me get through challenging times.

I would not be where I am today without my dearest friends, who stayed in Poland. My PhD adventure would not be possible without Dr. Agnieszka Graczyk-Jarzynka, who one day in November 2015 welcomed me in the Department of Immunology (Medical University of Warsaw) and became my first mentor. I am also grateful to Julia, Sara, Kasia and Ania, who, despite being thousands of kilometres away, were always keeping my spirits up.

All of this would not have been possible without the support of Cancer Research UK Cambridge Centre and Addenbrooke's Charitable Trust, who funded my research.

Finally, I would like to thank my family. I am grateful to my parents, brother and grandparents for believing in me and supporting even the most bizarre of my ideas, to Alek's mom and grandma for our weekly chats on FaceTime, and lastly to Alek for his endless patience and fabulous lemon tarts.

Contents

1 Introduction **15**

1.1 Molecular basics of RNA translation 16

1.1.1 Four stages of protein synthesis 18

1.1.2 Overview of ribosome biogenesis 23

1.2 Translational control 25

1.2.1 Regulation of the translation initiation 25

1.2.2 Regulation of translation elongation 27

1.2.3 Regulation of translation termination 28

1.3 Deregulation of translation in human cancers 29

1.3.1 Oncogenic and tumour suppressor pathways converge at controlling protein synthesis 29

1.3.2 Ribosome biogenesis and its oncogenic potential 32

1.3.3 Translation factors are frequently deregulated in cancer 33

1.3.4 The significance of translational response to stress in cancer 34

1.4 Elements of B-cell biology from the perspective of lymphoma development 38

1.5 Role of translation in B-cell development and malignancy 42

1.6 Toolkit to study heterogeneity of translation 46

1.7 Project aims 51

2 Materials and methods **53**

2.1 Materials 53

2.1.1 Overview of genomic sequences and annotations used in this study . 53

2.1.2 Overview of external datasets used in this study 54

2.1.3 Overview of computational software used in this study 55

2.2 Methods 56

2.2.1 Next Generation Sequencing library preparation and sequencing . . 56

2.2.1.1 RNA-Seq 56

2.2.1.2 Ribo-Seq 56

2.2.2 Processing and quality control of Next Generation Sequencing data 57

2.2.2.1 Adapter trimming and alignment to the reference genome 57

2.2.2.2 Read counting . . . 57
2.2.3 Differential translation analysis of Ribo-Seq data . . . 57
2.2.4 Differential expression and downstream analysis of RNA-Seq data . 58
2.2.5 Metagene analysis of iCLIP and Ribo-Seq data . . . 58
2.2.6 Differential expression analysis of RNA-Seq data . . . 58
2.2.7 Downstream data analysis . . . 58
2.2.7.1 Individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) . . . 59
2.2.8 Identification of DDX3X mutations from RNA-Seq data . . . 59
2.2.9 DLBCL Cell-of-origin identification from RNA-Seq data . . . 60
2.2.10 Chromosome Y expression identification from RNA-Seq data . . . 60
2.2.11 Hierarchical *de novo* identification of translated regions from Ribo-Seq data . . . 61
2.2.12 Reanalysis of published mass spectrometry datasets . . . 62
2.2.13 Analysis of proteogenomic data downloaded from OpenProt and sORFdb . . . 63
2.2.14 Evolutionary conservation of identified ORFs . . . 63
2.2.15 CRISPR screen design . . . 63
2.2.16 Figures preparation . . . 65
2.2.17 R and Bioconductor . . . 65

3 Genome wide quantification of translation in lymphoid malignancies 67
3.1 Background . . . 67
3.1.1 Establishing a bioinformatic pipeline for processing of translatome and transcriptome data . . . 69
3.1.2 Quality of translatome profiling in primary GC B-cells . . . 74
3.1.3 Benchmarking statistical approaches for differential translation analysis 79
3.2 Translational regulation following BCL6 and MYC overexpression in primary GC B-cells . . . 87
3.3 Discussion . . . 90

4 Mutations in RNA helicase DDX3X facilitate MYC-driven lymphomagenesis 95
4.1 Background . . . 95
4.2 Results . . . 98
4.2.1 Examining the prevalence and distribution of DDX3X mutations . . 98
4.2.1.1 DDX3X is preferentially mutated in MYC driven lymphomas 98
4.2.1.2 Context dependent pattern of DDX3X mutation in different cancer types . . . 101

4.2.1.3 DDX3X mutations in B-cell lymphomas cluster within C-terminal helicase domain . . . 101
4.2.1.4 Males with Burkitt Lymphoma and DLBCL are more likely to have DDX3X mutation . . . 103
4.2.2 DDX3X regulates ribosome biogenesis and global protein synthesis . 105
4.2.2.1 DDX3X binds preferentially to mRNA encoding components of core translation machinery . . . 105
4.2.2.2 DDX3X regulates translation of a subset of expressed transcripts . . . 110
4.2.3 Deregulation of MYC in primary GC B-cells increases ribosome biogenesis and triggers ER stress. . . . 118
4.2.4 DDX3X mutation interferes with endoplasmic reticulum stress response 120
4.2.4.1 DDX3X R475C mutation is associated with suppression of unfolded protein response in U2932 cells . . . 120
4.2.4.2 DDX3X mutation is associated with suppression of unfolded protein response in BL patients . . . 123
4.2.5 Up-regulation of DDX3Y in established tumours rescues loss of DDX3 helicase activity . . . 125
4.3 Discussion . . . 127

5 Elucidating the role of translated micropeptides in Diffuse Large B-cell Lymphoma 133
5.1 Background . . . 133
5.2 Results . . . 135
5.2.1 A systematic approach for *de novo* identification of noncanonical translation products in lymphoid cells. . . . 135
5.2.1.1 An integrated ORF identification workflow . . . 135
5.2.1.2 Pervasive translation of crude non-coding regions in lymphoid cells . . . 138
5.2.2 Noncanonical ORFs account for about 10% of proteins detected in proteomics experiments . . . 140
5.2.3 Characteristics of noncanonical ORFs producing MHC-bound peptides 144
5.2.4 Design of customised knockout CRISPR screen to identify noncanonical ORFs important for B-cells survival . . . 146
5.3 Discussion . . . 149

6 Perspectives 155

Bibliography 159

CHAPTER 1

Introduction

The 'central dogma' of molecular biology, proposed by Crick (1958), states that genes are chemically expressed as proteins in a sequence of molecular procedures. First, a DNA sequence is rewritten (*transcribed*) into RNA, then decoded (*translated*) into an amino-acids chain, which eventually folds into a protein **(Figure 1.1)**. For a long time, the translation process and its core components - the ribosomes and translation factors were viewed as molecular machines passively processing all available transcripts. This focus has been rewired by the ready avaliability of technology to quantify mRNA abundance, such as microarray or RNA-Seq.

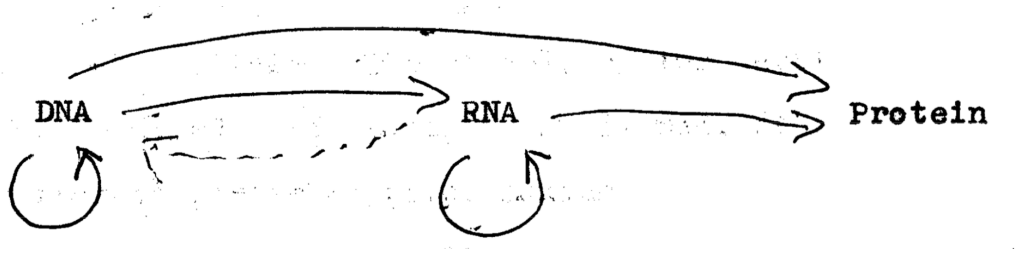

Figure 1.1: Francis Crick's first draft of the central dogma of molecular biology from an unpublished note (1956)

However, the relationship between mRNA and protein levels is far from being simple. Firstly, it is estimated that steady-state transcript levels in human cells can explain between 56% and 84% of protein abundance (Lundberg et al., 2010; Liu et al., 2016; Schwanhäusser et al., 2011; Jovanovic et al., 2015). The analysis performed by Jovanovic et al. (2015) showed that mRNA level, translation intensity and protein degradation rate can explain up to 79% of total protein abundance, with 18 - 26% and 8 - 22% of this corresponding to translation and protein degradation, respectively. Secondly, the relationship between mRNA and its product may vary when comparing different cell types and conditions. Relative changes in protein abundance during dynamic cell transitions are explained predominantly by the mRNA abundance, however, the rates of synthesis

and degradation vary substantially between individual proteins, which highlights the importance of regulation other than transcriptional (Jovanovic et al., 2015; Mathieson et al., 2018).

Therefore, there is a growing appreciation of the importance of regulation imposed at the point of translation. Protein synthesis consumes the majority of cellular energy resources, especially if a cell is bio-synthetically active, rapidly growing, or differentiating (Rolfe and Brown, 1997; Buttgereit and Brand, 1995; Lynch and Marinov, 2015). Hence, precise regulation of translation seems to be energetically efficient providing an opportunity to modulate cellular protein levels quickly. It is not surprising that translational control is especially important for highly energetic processes, such as cell proliferation, hormone release and stress response (Hershey et al., 2012). Translation is also the primary mechanism of gene expression regulation in cells lacking active transcription (for example, in oocytes or red blood cells) or during the early stages of viral infection, when the host transcription is suppressed (Hershey et al., 2012; Mohr, 2016). Deregulated translation is a hallmark of cancer, but the exact role of this is complex and context dependent.

1.1 Molecular basics of RNA translation

Translation is an evolutionarily conserved process during which the nucleotide sequence of a messenger RNA (mRNA) is decoded into the amino-acid sequence of a protein by rules known as genetic code. A CoDing Sequence region (CDS) is a region of mRNA or genomic DNA whose nucleotide sequence determines amino-acids sequence. It is usually flanked by two UnTranslated Regions (UTRs): 5′UTR and 3′UTR. Every three adjacent nucleotides of a CDS form a unit called a codon. There are 4 x 4 x 3 = 64 possible combinations of the three nucleotides, three of which indicate the end of a protein; the remaining 61 corresponds to one out of 20 amino acids. An Open Reading Frame (ORF) is a series of codons contained between a start and a stop codon. A given RNA sequence can be decoded in one out of three possible reading frames depending on position of the first (start) codon.

The codons of a protein coding sequence are deciphered from 5′ end to 3′ end by transfer RNA (tRNA). tRNAs are small adaptor molecules that pair with a coding triplet through a complementary anticodon. The genetic code is redundant, which means that more than one codon specifies a single amino acid, and more than one tRNA molecule can match a single codon. Some tRNAs require a complete match between codon and anticodon, but for some, one mismatch (wobble) at the third position of the codon does not affect the accurate amino-acid determination. A tRNA carrying a matching amino acid is called aminoacyl-tRNAs.

The majority of protein coding mRNAs undergoe two specific post-transcriptional

modifications: polyadenylation of the 3′ end and capping of the 5′end with a methylated guanosine at 7th position (m^7G).

The actual reactions of protein synthesis take place inside the ribosome, a large ribonucleoprotein complex (RNP) composed of multiple ribosomal proteins and ribosomal RNAs (rRNA). A fully assembled eukaryotic ribosome (80S) consists of a small (40S) and a large subunit (60S). When not involved in active translation, the ribosomal subunits are separated. During evolution, the ribosome became larger and richer, but its core components are so highly conserved that a common ancestor of all living species must have synthesised its proteins similarly to organisms living today. The human ribosome contains 80 ribosomal proteins and four separate rRNA molecules that account for more than 80 % of the cellular RNA pool (Cech, 2000). Each ribosome has three tRNA binding sites, that is A-site for **a**minoacyl-tRNA, P-site for **p**eptidyl-tRNA and E-site for **e**xit **(Figure 1.2)**. Each covers three adjacent codons and is essential for the orchestration of the ribosome movement along the coding sequence. The ribosome reads only one codon at a time, but several ribosomes (polysomes) can simultaneously translate a single mRNA molecule.

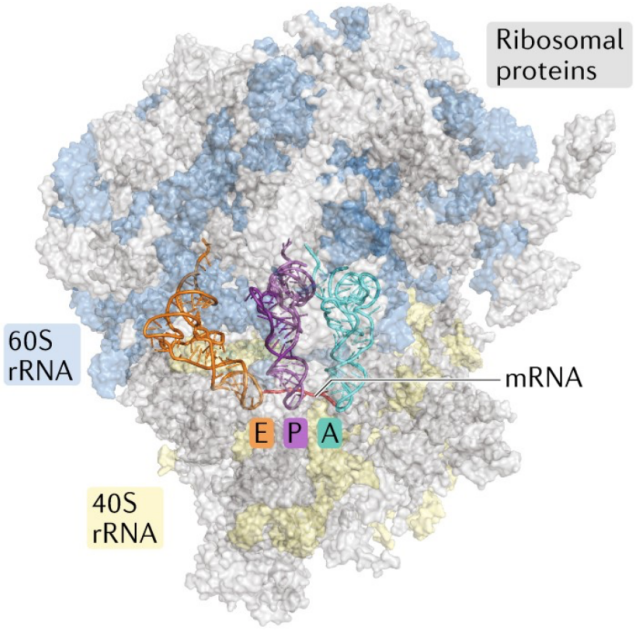

Figure 1.2: Structural model of the eukaryotic 80S ribosome
Three sites for tRNA binding are highlighted: E-site - orange, P-site - violet and A-site - green. mRNA is shown in red. Estimated localisation of the ribosomal proteins is indicated by shaded areas: blue for 60S and yellow for 40S subunit; from Schuller and Green (2018)

1.1.1 Four stages of protein synthesis

Translation process consists of four phases: initiation, elongation, termination and recycling.

Translation initiation

Translation initiation is a multi-step process, which is considered the principal point of translational control. There are two main modes of translation initiation: 5′cap dependent and 5′cap independent, of which 5′cap dependent is the most common.

The canonical pathway of 5′cap dependent translation initiation can be divided into 4 stages. The ternary complex (TC), which consists of methionine carrying tRNA (Met-tRNA or methionyl-tRNA), GTP and eukaryotic Initiation Factor 2 (eIF2) complex is loaded to the small ribosomal subunit (40S) recycled from the previous round of translation with bound eIF1, eIF1A and eIF3 factors. TC together with 40S, eIF1, eIF1A and eIF3 form a 43S preinitiation complex (43S PIC) **(Figure 1.3, steps 1-2)** (Jackson et al., 2010). Recruitment of 43S PIC onto mRNA requires cooperation of eIF4F complex, eIF4B and eIF4H. eIF4F complex, consists of a 5′cap-binding eIF4E, RNA-helicase eIF4A and a scaffold protein eIF4G **(Figure 1.3, step 3)**. Apart from interacting with eIF4A and eIF4E, eIF4G binds also poly(A) binding protein (PABP), which brings mRNA into circular shape with 5′cap and 3′-poly(A) tail close together creating the so-called *closed-loop* structure (Mangus et al., 2003; Taylor et al., 2020). 43S PIC starts base-by-base scanning of 5′UTR. It consists of coordinated unwinding of the 5′UTR secondary structures and 43S PIC movement towards 3′end. Once 43S PIC recognises a start codon (usually AUG encoding methionine), the scanning process stops (Jackson et al., 2010). Sometimes an optimal start codon is not necessarily the first AUG encountered. In most eukaryotic mRNAs, a start codon is framed into a Kozak consensus sequence, an RNA sequence context ensuring the fidelity of the translation initiation site (TIS) (Kozak, 1987). If the context of a potential start codon is weak (poorly resembling a Kozak sequence), PIC may skip it and continue to search for the next codon. This phenomenon is known as *leaky scanning* (Zhang et al., 2019). The selection of the optimal TIS is promoted by eIF1, which discriminates between AUG and weak context codons. The commitment to a TIS is promoted by eIF5, a GTPase-activating protein targeting eIF2 complex and leading to partial dissociation of eIF2 from 40S subunit (Jackson et al., 2010). Assembly of 80S ribosome at the TIS is mediated by eIF5B and followed by release of eIF1, eIF1A, eIF3 and residual eIF2. **(Figure 1.3, step 4)**. The ribosome is now ready to enter the elongation stage (Aitken and Lorsch, 2012; Hinnebusch, 2014).

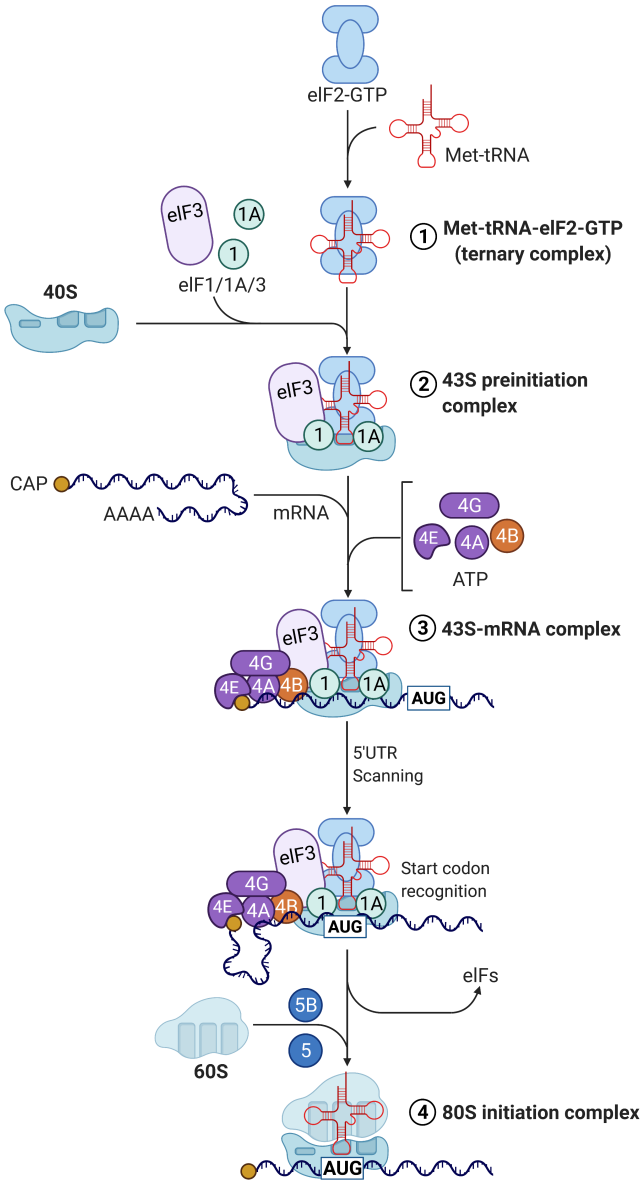

Figure 1.3: Diagram showing a simplified process of 5′cap- dependent translation initiation in eukaryotic cells.

The process of translation initiation is divided into four steps:

Step 1. Ternary complex formation from GTP-bound eIF2 and methionyl-tRNA (Met-tRNA).

Step 2. 43S preinitiation complex (43S PIC) assembly, which includes the ternary complex, the small ribosome subunit 40S and translation initiation factors (eIF3, eIF1A and eIF1)

Step 3. mRNA binding to 43S PIC promoted by eIF4F complex and eIF4B followed by 5′UTR scanning in 5′ to 3′end direction. An optimal translation initiation site, typically AUG codon framed by Kozak sequence, is promoted by eIF1 and eIF5

Step 4. Recruitment of the large ribosome subunit 60S leading to 80S ribosome assembly, and dissociation of the initiation factors mediated by eIF5B.

Adapted from Protein Translation Cascade, by BioRender.com (2021). Retrieved from `https://app.biorender.com/biorender-templates`

A less common mechanism of translation initiation involves 5′cap-independent ribosome recruitment. Although the exact mechanisms of cap-independent translation are not fully understood, the two proposed mechanisms involve Internal Ribosome Entry Sites (IRES) and 5′cap-independent translational enhancers (CITEs). Translation through IRES is a common strategy employed by pathogenic viruses to escape global halt of 5′cap mediated translation in the host cell, but was also found in eukaryotic genes, particularly those involved in stress or anti-viral response (Yang and Wang, 2019; Jackson et al., 2010). IRES can interact with canonical initiation factors and recruits 40S ribosomal subunit through IRES trans-acting factors (ITAFs) (Komar and Hatzoglou, 2011; Yang et al., 2017; Meyer et al., 2015) or short cis-elements that pairs with 18S rRNA (Dresios et al., 2006; Yang and Wang, 2019). Because of the lack of conserved IRES sequence, the exact number of IRES-initiating open reading frames and factors regulating this mode of translation is unknown.

An alternative mechanism of 5′cap-independent translation, also originating from viruses, involves CITEs - RNA structural elements within mRNAs attracting translation initiation factors (Shatsky et al., 2018). Proposed mechanism of CITEs-mediated translation initiation involves reversible N^6A-methylation of mRNA within GAC sequence context recognised by YTHDF1 and followed by direct recruitment of the eIF3 complex (Meyer et al., 2015; Wang et al., 2015; Shatsky et al., 2018). Although it is still a widely debated topic, both CITEs and IRES initiating translation may proceed without a full set of translation initiation factors (Terenin et al., 2013).

It is important to highlight that 5′cap-independent translation is still a largely unexplored territory and some controversies arose around the accuracy of IRES reporter assays and so the existence of IRES in eukaryotic cells (Shatsky et al., 2018; Yang and Wang, 2019). Nevertheless the concept of alternative translation initiation adds another layer of complexity to the cellular translatome potentially broadening proteome diversity.

Translation elongation

Translation elongation takes place between loading the first aminoacyl-tRNA after the start codon and encountering the first in-frame stop codon. Translation proceeds in the 5′ - to - 3′ direction, so the N-terminal end of a protein is translated first (Schuller and Green, 2018). The elongation stage consists of three major steps: tRNA binding, peptide bond formation and translocation.

After initiation is completed, the tRNA carrying the first N-terminal amino-acid (usually Met-tRNA) is localised in the middle slot of the ribosome (P-site). Next slot (A site) exposes the next codon. Once tRNA carrying the next amino-acid binds to the complementary A site, eukaryotic Elongation Factors 1 alpha 1 (eEF1A1) hydrolyses GTP to GDP fixing aminoacyl-tRNA within the A-site. The next stage is the formation of a

peptide bond between two adjacent amino-acids. The peptidyl transferase centre of the large ribosomal subunit catalyses the reaction, after which the aminoacyl-tRNA occupying the P-site releases the attached amino-acid. The ribosome is now in the pre-translocation state with a peptidyl-tRNA (tRNA with nascent peptide attached) in the A-site and deacetylated tRNA (tRNA with detached amino-acid) in the P-site (Schuller and Green, 2018). The last step involves the a series of conformational changes that pushes the entire ribosome three nucleotides towards the 3' end - a reaction conjugated with another GTP hydrolysis reaction catalysed by eEF2. During the translocation phase, deacetylated tRNA from the P-site is transferred to the E-site and released (Ratje et al., 2010). The A-site is now empty and ready to accept a new aminoacyl-tRNA. The elongation cycle then repeats until the termination codon appears in the A-site. Each elongation cycle will add one new amino acid to the growing polypeptide chain **(Figure 1.4)**.

Efficient conformational changes of the ribosome during each elongation cycle are essential for maintaining the right direction of translation. The energy for this process is delivered by GTP to GDP hydrolysis, which is performed by the two eucariotic Elongation Factors 1 and 2 (eEF1 and eEF2).

The accuracy of the ribosome in decoding mRNA into protein is estimated to reach almost 99,99% (1 misincorporation per 10^4 amino acids joined). Because the release of faulty protein product could potentially cause serious consequences for the cell, two proofreading mechanisms monitor each elongation cycle (Neelagandan et al., 2020). Correct codon-anticodon matching is favoured due to its high affinity, which stabilises the bond formation, affects rRNA folding around the tRNA-rRNA interaction and triggers GTP hydrolysis by eEF1. When this is missing, the elongation becomes slow, aminoacyl tRNA cannot be fixed in the A-site and dissociates off the ribosome. An invalid tRNA in the P-site increases the risk for further decoding errors, thus decreasing chances for synthesising a full-length protein. Repetitive amino-acid misincorporation may lead to premature termination of translation (Neelagandan et al., 2020).

Translation termination

The end of the coding sequence is marked by a stop codon (UAA, UAG or UGA). All three are recognised by eukaryotic release factor 1 (eRF1). Another release factor, eRF3, is a GTP-ase that promotes the hydrolysis of the peptidyl-tRNA bond in the P-site. This releases the C-terminal end of the newly synthesised amino-acid chain from the ribosome. The nascent polypeptide then completes folding in the cytoplasm. The post-termination complex (post-TC) disassembles and can be recycled to participate in another round of translation (Hellen, 2018).

Ribosome recycling

The ribosome recycling stage aims to split 80S ribosomes into separate subunits, preparing them for a new round of translation. This is initiated by the recruitment of ABCE1 to post-termination ribosomes with eRF1 in the A-site. After 80S disassembly, the last step involves releasing deacetylated tRNA and mRNA from the 40S ribosome subunit, which is mediated by translation initiation factors: eIF1, eIF1A and eIF3. The full process of translation termination and recycling was reviewed comprehensively by Hellen (2018)

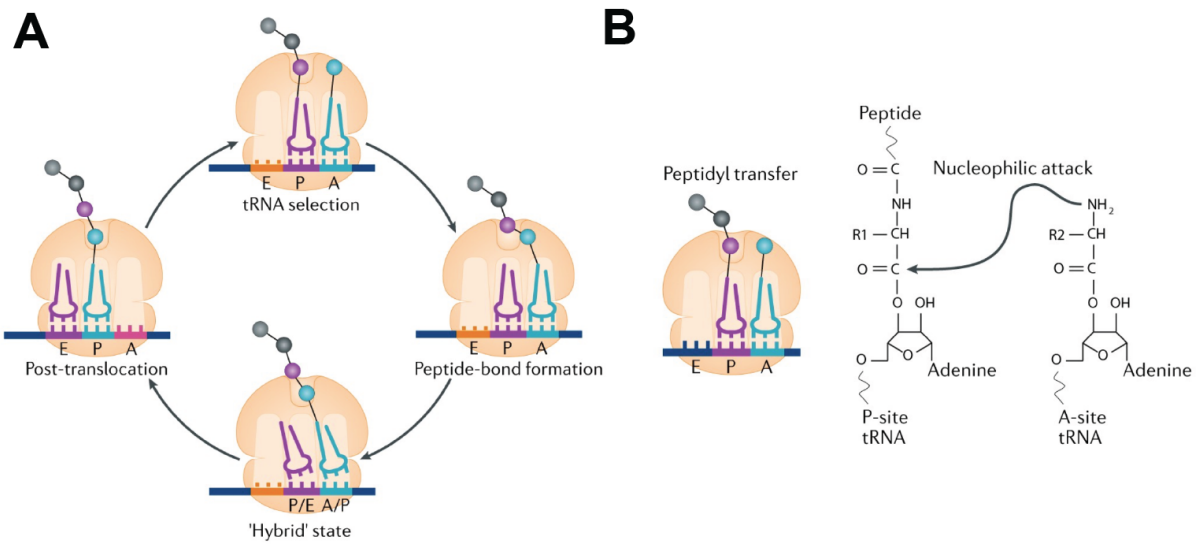

Figure 1.4: Diagram showing the process of translation elongation in eukaryotic cells.

A) Schematic of the translation elongation cycle demonstrating how tRNA moves between the ribosomal sites:
Amino-acyl tRNA (green) recognises codon in the ribosomal A-site. This is followed by peptide-bond formation and transfer of the nascent peptide chain to amino-acyl tRNA in the A-site, now a peptidyl-tRNA. Change in the ribosome conformation after peptide bound formation are referred to as a *hybrid state* - a transient conformation, where anticodon loop of the tRNA remains fixed in in the P and A sites of the small ribosomal subunit, but the amino-acid site of tRNA (the acceptor stem) is already in the E and P sites of the large ribosomal subunit (Schuller and Green, 2018) The process of the ribosome translocation ends with peptidyl-tRNA in the P-site and A-site ready to accept the next amino-acyl tRNA (Schuller and Green, 2018).

B) Overview of the peptide-transfer reaction catalysed by the ribosome. The peptide-bond formation occurs through nucleophilic attack of the amino group of the new amino acid (bound to tRNA in the A-site) on the ester linkage on the peptidyl-tRNA (remaining in the P-site) (Schuller and Green, 2018). This leaves a deacetylated tRNA in P-site and peptidyl-tRNA in the A-site longer by one amino-acid. From Schuller and Green (2018)

1.1.2 Overview of ribosome biogenesis

Ribosome biogenesis is a complex process involving biosynthesis and assembly of the complete 80S human ribosome. Briefly, 80S ribosome consists of 2 subunits: small and large, both composed from ribosomal RNA (rRNA) and ribosomal proteins (RPs). The small subunit (40S) is responsible for binding, scanning and unwinding of the mRNA, while the function of the large subunit (60S) is to catalyse peptide bond formation and check the quality of the nascent peptide (Pelletier et al., 2018).

The nucleolus, a nuclear substructure, is central to the process of ribosome biogenesis. It is responsible for transcription, processing and modifications of rRNA and the assembly of precursor ribosomal subunits (Lafontaine et al., 2021). rRNA sequences are organised in clusters of tandem repeats encoded by ribosomal DNA (rDNA) in nucleolus organiser regions (NORs). NORs contain 18S, 5.8S and 28S rRNA sequences (47S pre-rRNA) separated by spacers with regulatory sequences and distributed between the short arms of five acrocentric chromosomes 13, 14, 15, 21 and 22 (Henderson et al., 1972; Pelletier et al., 2018). 47S pre-rRNA is transcribed by RNA polymerase I. Another ribosome component, 5S rDNA cluster, is localised on chromosome 1, outside the nucleolus, and transcribed by RNA polymerase III (Pol III). In contrast, mRNAs of RPs are transcribed by RNA polymerase II (Pol II), exported to the cytoplasm for translation and then re-imported to the nucleus to participate in the ribosome assembly (Pelletier et al., 2018). In the nucleolus, 47S pre-rRNA, 5S rRNA, numerous RPs and assembly factors co-transcriptionally form a 90S processome (pre-ribosome) **(Figure 1.5)**.

Next step is 90S maturation, which includes extensive base modifications and cleavage reactions, resulting in the separation of the pre-40S and pre-60S subunits (Pelletier et al., 2018; Lafontaine, 2015). This requires the activity of small nucleolar RNAs (snoRNAs) derived from introns of certain Pol II transcribed genes. Finally, the subunits are exported to the cytoplasm for the final maturation (incorporation of few additional RPs and accessory factors) contributing to the assembly of 80S ribosome for protein synthesis **(Figure 1.5)** (Tschochner and Hurt, 2003; Pelletier et al., 2018).

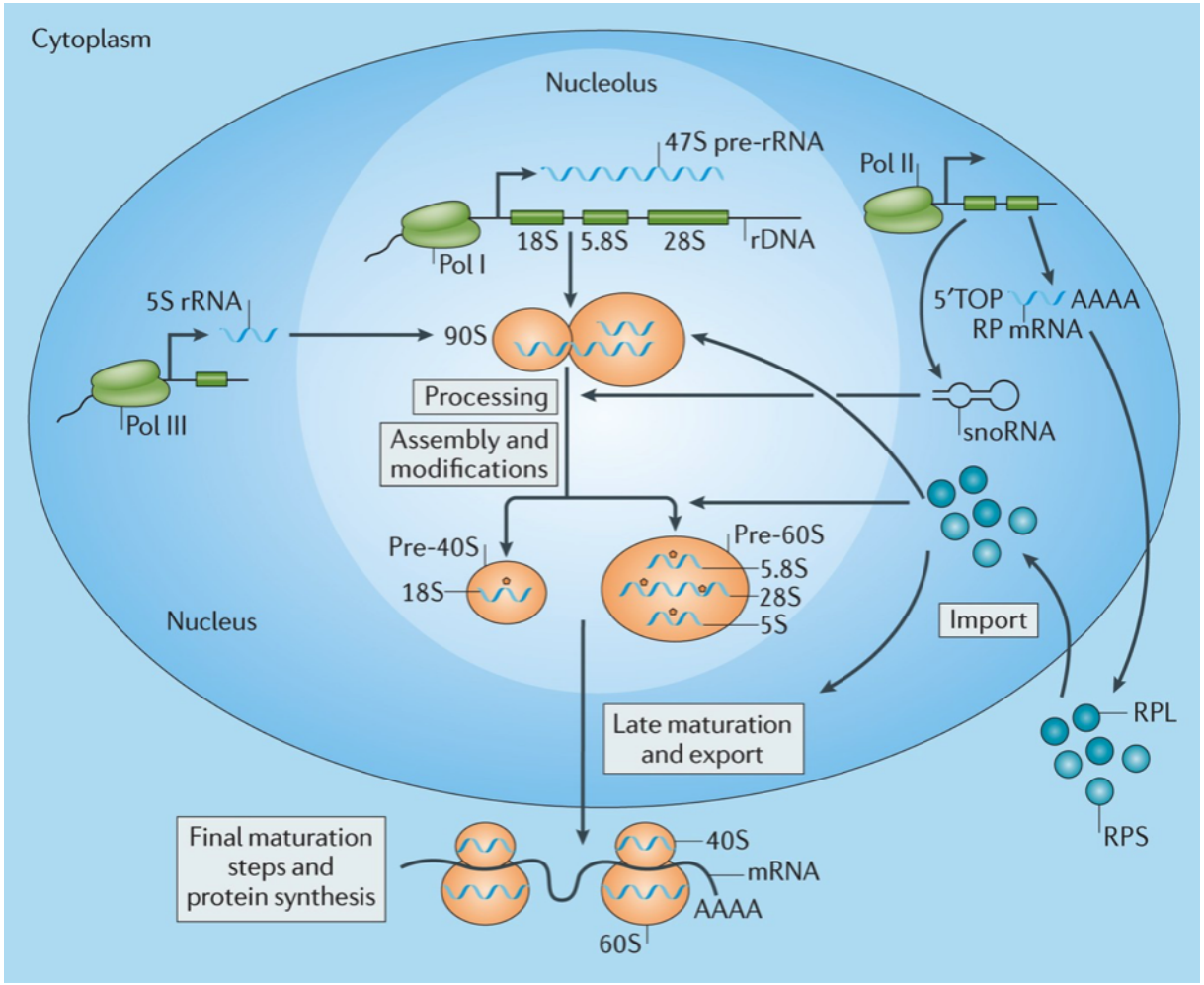

Figure 1.5: Overview ribosome biogenesis

The components of eukaryotic ribosomes are the product of three RNA polymerases transcribing different parts of ribosomal RNA (rRNA), and mRNA for ribosomal proteins (RP). Polymerase III (Pol III) is responsible for transcription of 5S rRNA cluster in the nucleus, while polymerase I (Pol I) in nucleolus produces 47S pre-RNA containing 18S, 5.8S and 28S rRNA. As is typical for other protein-coding regions, the RP sequence is transcribed by Pol II in the nucleus. RP's mRNA is then exported to the cytoplasm for translation and imported back to the nucleus. RPs, 47S pre-RNA and 5S rRNA participate in the assembly of 90S processome, which undergoes chemical modifications and cleavage resulting in formation of two separate ribosomal subunits (pre-40S and pre 60S). Pre-40S and pre 60S are exported to the cytoplasm, where after final maturation step, they are ready to participate in protein synthesis. From Pelletier et al. (2018)

1.2 Translational control

Regulation of translation is a broad term covering mechanisms affecting different stages of protein synthesis. A given mechanism may affect translation globally or be specific towards a single mRNA or a transcript group. A further distinction can be made between processes that modify the function of the core components of the translation machinery and those that affect RNA directly. From a molecular perspective, translational control includes a variety of mechanisms such as chemical modifications (e.g. phosphorylation or methylation), differential expression of translation factors, modulation of the 3D structure of mRNA, action of trans-activating RNA-binding proteins (RBPs), presence of cis-regulatory elements in mRNA or recruitment of microRNAs to 3′UTR.

For clarity, I will review the most important points of regulation, focusing on each stage of translation separately.

1.2.1 Regulation of the translation initiation

Translation initiation is considered the main rate-limiting step of eukaryotic protein synthesis. Many of the regulatory mechanisms target the balance between 5′cap dependent and 5′cap independent translation by modulating the activity of core translation initiation factors.

The most extensively studied mechanisms of translation initiation control involves the regulation of eIF4E activity through 1) transcription, 2) phosphorylation or 3) sequestration by binding to a family of translational repressors (Raught and Gingras, 1999). eIF4E is a direct transcriptional target of many important signalling pathways, including NF-κB and c-MYC (Hariri et al., 2013). Phosphorylation of eIF4E at Ser 209 is mediated by MAP kinase-interacting kinases MNK1/MNK2 and occurs in response to various mitogenic and stress factors promoting 5′cap dependent translation initiation (Sonenberg, 1996; Pyronnet et al., 1999). Another arm of regulation is formed by the family of eIF4E binding proteins (EIF4EBPs), which bind eIF4E and remove it from the pool available for translation. Binding of 4E-BPs and eIF4G, a part of eIF4F complex, are mutually exclusive for eIF4E. (Yang et al., 2020). A well characterised regulator of this process is mTOR signalling, which couples the activity of growth factors with nutrients availability and the intensity of anabolic processes (Saxton and Sabatini, 2017; Liu and Sabatini, 2020; Kim and Guan, 2019a). mTOR-mediated phosphorylation of EIF4EBPs disrupts the 4E-BP–eIF4E complex allowing eIF4E to participate in eIF4F initiation complex. EIF4E is required for translation of all 5′capped mRNAs, but not all transcripts are equally susceptible for changes in eIF4E levels. Transcripts with long 5′UTRs and specific RNA regulatory motifs are particularly sensitive to eIF4E levels. (Smith et al., 2021).

The 5′UTR scanning requires coordinated unwinding of mRNA structure, which is performed by RNA helicases, typically eIF4A - a component of the eIF4F complex. eIF4A activity is promoted by two cofactors: eIF4B or eIF4H, while the avaliability by programmed cell death 4 (PDCD4). (Yang et al., 2003; Smith et al., 2021). PDCD4 acts downstream of mTOR and is controlled by inactivation phosphorylation by Ribosomal Protein S6 Kinase (S6K) (Silvera et al., 2010). Although the net effect of eIF4A inhibition or sequestration is a mild reduction in global translation rate, the rate and even the direction of change vary between single transcripts (Modelska et al., 2015; Smith et al., 2021). eIF4A-sensitive mRNAs are characterised by complex secondary structures in their 5′UTRs and the presence of specific RNA motifs (Steinhardt et al., 2014; Modelska et al., 2015; Wolfe et al., 2014a).

Another important mechanisms of translation initiation control involves phosphorylation of eIF2α, the main regulatory subunit of eIF2 complex, which is responsible for joining methionine-tRNA to the small ribosomal subunit (Smith et al., 2021). Phosphorylated eIF2α has an inhibitory effect on global translation, because it prevents GDP to GTP restoration of eIF2β, another eIF2 subunit, trapping it its inactive form. This shifts the balance towards 5′cap-independent translation initiation. Translation control through eIF2α is central to the Integrated Stress Response (ISR), an adaptive pathway, which aims to restore cellular homeostatsis or commit the cell to apoptosis following exposure to unfavourable conditions (Pakos-Zebrucka et al., 2016), see section 1.3.4.

The efficiency of translation initiation can also be controlled by numerous RNA-binding proteins, of which RNA helicases deserves special attention. RNA helicases are multifunctional proteins involved in all aspects of RNA metabolism, such as mRNA processing, nuclear export, trafficking in the cytoplasm, translation, degradation or microRNAs-mediated RNA silencing. The activity of RNA helicases regulates almost all stages of protein synthesis (Bourgeois et al., 2016; Linder and Jankowsky, 2011). The efficiency of 5′cap-dependent translation initiation relies on the ability to unwind structured 5′UTRs. Helicases assisting in this process include DDX48, DDX3X and components of eIF4F complex (eIF4A1 and eIF4A2). Those are especially important for translating mRNAs with GC-rich 5′UTRs and 3′UTRs with microRNA binding sites (Bourgeois et al., 2016; Linder and Jankowsky, 2011; Sen et al., 2015; Wolfe et al., 2014b). Other roles of RNA helicases in translation control include mRNA positioning at ribosomal 40S subunit (DHX29), regulation of 40S scanning and ribosome recycling (DHX9), promoting 80S ribosome assembly (DHX33, DDX3X) and recognition of stop codon (DDX19B). RNA helicases also control the translation of mRNA regulons; for example, DDX25 is essential for the translation of mRNAs associated with spermatogenesis, and DDX48 facilitates the translation of several neuronal mRNAs (Bourgeois et al., 2016; Linder and Jankowsky, 2011). Despite recent

advances in the biology of RNA helicases, the full spectrum of their activity remains elusive.

tRNA-derived small RNAs (tsRNAs), a novel class of regulatory non-coding RNAs, compose another interesting layer of translational control. tsRNA are cleaved from several types of tRNA in response to stress. These tsRNAs contains an oligo-G terminal motif forming G-quadruplexes in the 5′UTR of certain transcripts, that suppress translation initiation by displacing eIF4F complex from 5′cap (Ivanov et al., 2014). This mechanism of translation inhibition by stress-induced tsRNAs was identified in several organisms and cell types indicating high evolutionary conservation (Ivanov et al., 2011).

1.2.2 Regulation of translation elongation

For a very long time translation elongation was not considered an important rate-limiting step of protein production. However, variety of factors can influence the elongation speed, which is especially important in the terms of protein folding and transport. Regulatory mechanisms of translation elongation involve: ORF codon composition, local sequence context, post-transcriptional modifications of tRNA, mRNA, and the expression levels of the core elongation factors (Knight et al., 2020; Schuller and Green, 2018).

The genetic code is redundant (degenerate), so each amino acid can be decoded by two, four or six codons, with exception for methionine and tryptophan, which are encoded by a unique codon. Interestingly, the distribution of synonymous codons over the transcriptome is not uniform (codon bias), and the rate of translation of different codons is uneven (codon optimality). Hypotheses put forward to explain an evolutionary origin of such phenomenon involve varying elongation rate of different codons, translation accuracy or selection of splicing enhancers (Richter and Coller, 2015). The speed of tRNA-codon pairing depends on the abundance of a specific tRNA in the cytoplasm (it takes more time for a less abundant tRNA to find the codon) and interaction strength (standard Watson-Crick pairing is quicker than wobble pairing). The relative abundance of different types of tRNA in human is not uniform and varies between tissues promoting translation of cell-specific transcripts (Dittmar et al., 2006). This view has been expanded by findings in yeast, that demonstrated coordinated changes in the abundance of specific tRNAs after exposure to various types of stress. In consequence, when tRNA pool is restricted, transcripts with larger number of *rare* codons tends to be translated less efficiently (Torrent et al., 2018; Hanson and Coller, 2018). Codon bias correlates with tRNA levels, therefore, to some extent, it is mirrored by codon optimality (Sabi and Tuller, 2014; Knight et al., 2020). Both translation efficiency and mRNA half-life correlate with codon composition, but this changes with the concentration of specific tRNAs in the cytoplasm (Presnyak et al., 2015). Although variations in codon selection and translation dynamics have been

observed within and between all species, the majority of studies in this topic came from lower organisms (Hanson and Coller, 2018).

Certain codon patterns, for example poly-lysine tracks, or mRNA secondary structures can lead to programmed ribosome stalling and frameshifting (Schuller and Green, 2018). Local slowdown, or even halt in the elongation rate may facilitate protein folding or signal recognition particle binding, which essential for secreted protein (Richter et al., 2012). Ribosome transit can be regulated by several protein interacting with the A-site, for example the elongation factor EF-G. Programmed frameshifting involves either slipping back or skipping one nucleotide and occurs in all species (Ketteler, 2012). The scope of this phenomenon in human is still debated. An example of a gene known to regulate its translation through frameshifting is ornithine decarboxylase (ODC) (Bekaert et al., 2008).

1.2.3 Regulation of translation termination

Although termination of translation is usually not a rate-limiting step of protein synthesis, the fidelity of this process is important for maintaining proteome integrity. The biological activity of truncated or extended proteins, possible products of noncanonical termination of translation, may be different from the original mRNA product.

Nonsense mediated decay (NMD) is a translation-coupled process of elimination of mRNAs harbouring premature stop codons. This is an important surveillance mechanisms preventing the expression of truncated proteins (Kurosaki and Maquat, 2016; Hellen, 2018). The role of NMD is not limited to aberrant transcript elimination. Transcripts with upstream ORFs in their 5′UTR, products of alternative splicing, intron retention or auto-regulatory loops may utilise programmed NMD to regulate the pool of mRNAs available for translation (Kurosaki and Maquat, 2016). It is estimated that the expression level of about 10% of eukaryotic mRNAs may be modulated by NMD (Kurosaki and Maquat, 2016)

Finally, near cognate recognition of stop codon (readthrough) may lead to an extended product, but the full scope of this phenomenon in mammalian cells is still unclear. Functional readthrough has been shown for three human genes: VEGFA, LDHB, and MDH1 (Schueren and Thoms, 2016). Interestingly, the extended isoforms show different subcellular location (LDHB, MHD1) or have the opposite biological activity to a canonical product (antiangiogenic instead of proangiogenic in case of VEGFA) (Eswarappa et al., 2014; Schueren and Thoms, 2016).

1.3 Deregulation of translation in human cancers

Enlarged nucleoli, which are the primary site of ribosome biogenesis, was one of the first hallmarks of malignancy that was later widely applied in diagnostics (Pianese, 1896; Gani, 1976). Deregulation of translation machinery is frequently observed in spontaneous cancers as well as in hereditary cancer syndromes. While the latter usually involves point mutations affecting ribosome biogenesis, the scope of translational reprogramming in the majority of cancers encompass a variety of mechanisms. Protein synthesis can be hijacked by malignant cells through aberrant expression, point mutations or post-translational modifications of translation factors or their regulators, mRNA regulatory elements, RNA modifications, or preferential codon usage. The mechanisms can include global changes in protein synthesis and/or increase or decrease in translation intensity of a subset of transcripts.

1.3.1 Oncogenic and tumour suppressor pathways converge at controlling protein synthesis

At the global scale, increased protein synthesis rate was shown to accompany high mitotic activity (Johnson et al., 1976). Indeed, many key oncogenic and tumour suppressor pathways, such as MYC, PI3K, RAS, PTEN, TP53, converge at the regulation of cellular translation synchronising proliferation rate with anabolic and catabolic pathways, but this relationship is far from being simple.

Deregulation of MYC is observed in >50% of human cancers (Meyer and Penn, 2008). The oncoprotein MYC family consists of three genes: *MYC*, *MYCN*, and *MYCL*, which have the capacity to regulate about 15% of human genes (Dang, 2012) coordinating a variety of cellular functions including proliferation, metabolism, differentiation and immunosurveillance (Chen et al., 2018). MYC regulates protein synthesis through controlling the transcription of ribosomal proteins (Boon et al., 2001), ribosomal RNA (rRNA) (Grandori et al., 2005) **(Figure 1.6)**, and translation initiation factors (eIF4A, eIF4E, and eIF4G) (Schmidt, 2004; Xu and Ruggero, 2020). Although the role of MYC far exceeds the regulation of translation, its oncogenic potential is highly dependent on the translation apparatus: the cell ability to phosphorylate eIF4E (Pourdehnad et al., 2013) and augment protein synthesis (Barna et al., 2008), see section 1.5. Moreover, MYC protein abundance is regulated at the level of translation. Firstly, through alternative, IRES-mediated, translation initiation from a CUG start codon and secondly, by preferential translation controlled by eIF4A and eIF4E (Schatz et al., 2011; Culjkovic-Kraljacic et al., 2016).

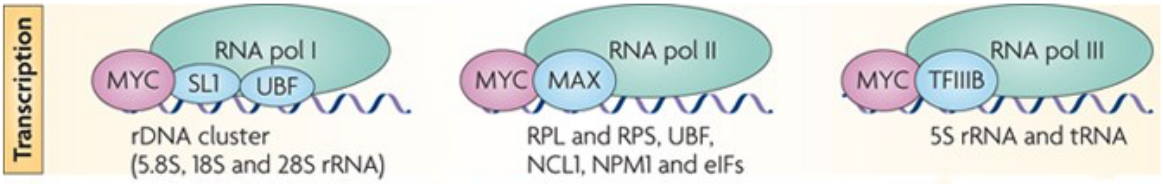

Figure 1.6: Multilevel regulation of ribosome biogenesis by MYC
MYC controls ribosome biogenesis by promoting transcription of ribosome biogenesis components. It does so by cooperation with other cofactors regulating the recruitment of RNA polymerases (RNA pol I, II and III) and through chromatin structure remodelling (Shiue et al., 2009; Van Riggelen et al., 2010). Adapted from Van Riggelen et al. (2010)

Interestingly, some ribosomal proteins were found to inhibit MYC expression creating a negative feedback loop. For example, RPL11 interferes with MYC binding to 5S rRNA and tRNA promoters (Dai et al., 2010) and, together with RPL5, jointly binds to MYC mRNA directing it to RNA-induced silencing complex (RISC) for degradation (Liao et al., 2014).

The translational control of PTEN/PI3K/AKT and RAS/MEK1/ERK signalling converges at regulating the activity of mTORC1 complex, a key coordinator of cell growth, survival and metabolism **(Figure 1.7)** (Kim and Guan, 2019b). Activation of these pathways leads to inactivating phosphorylation of TSC1/2 complex, which is a negative regulator of mTORC1. The mTOR pathway promotes 5′cap-dependent translation by controlling eIF4E inhibitory binding proteins, which sequester eIF4E and prevent it from participating in translation initiation (Silvera et al., 2010). Additional control is imposed at the level of mTORC1 targets. MNK1 and MNK2, which are downstream of RAS/MEK1/ERK, phosphorylate eIF4E at a single residue Ser209 independently of mTORC1 control, thus selectively increasing translation intensity of several protumorigenic transcripts in mice and human cell lines including anti-apoptotic MCL1 and cyclin D1 (CCND1), MYC, and proangiogenic VEGF and FGF2 (Furic et al., 2010; Sonenberg and Hinnebusch, 2009; Kevil et al., 1996). Biophysical studies demonstrated that phosphorylation at Ser209 promotes 5′cap independent translation, but the significance of this for tumourgenesis is unclear (Zuberek et al., 2003). We know that oncogenic capacity of both RAS/MEK1/ERK and MYC signalling depends on the ability to phosporylate eIF4E (*in vitro* and *in vivo* evidence) (Yang et al., 2020) suggesting that a specific, prooncogenic, mode of protein synthesis is a common strategy during cell transformation. Prooncogenic translation is not limited to malignant cells, though. Robichaud et al. (2018) showed that mice with mutated phosphorylation site (eIF4E^{S209A}) are resistant to the development of lung metastases due to decreased translation of MCL1 and BCL2 in prometastatic neutrophils.

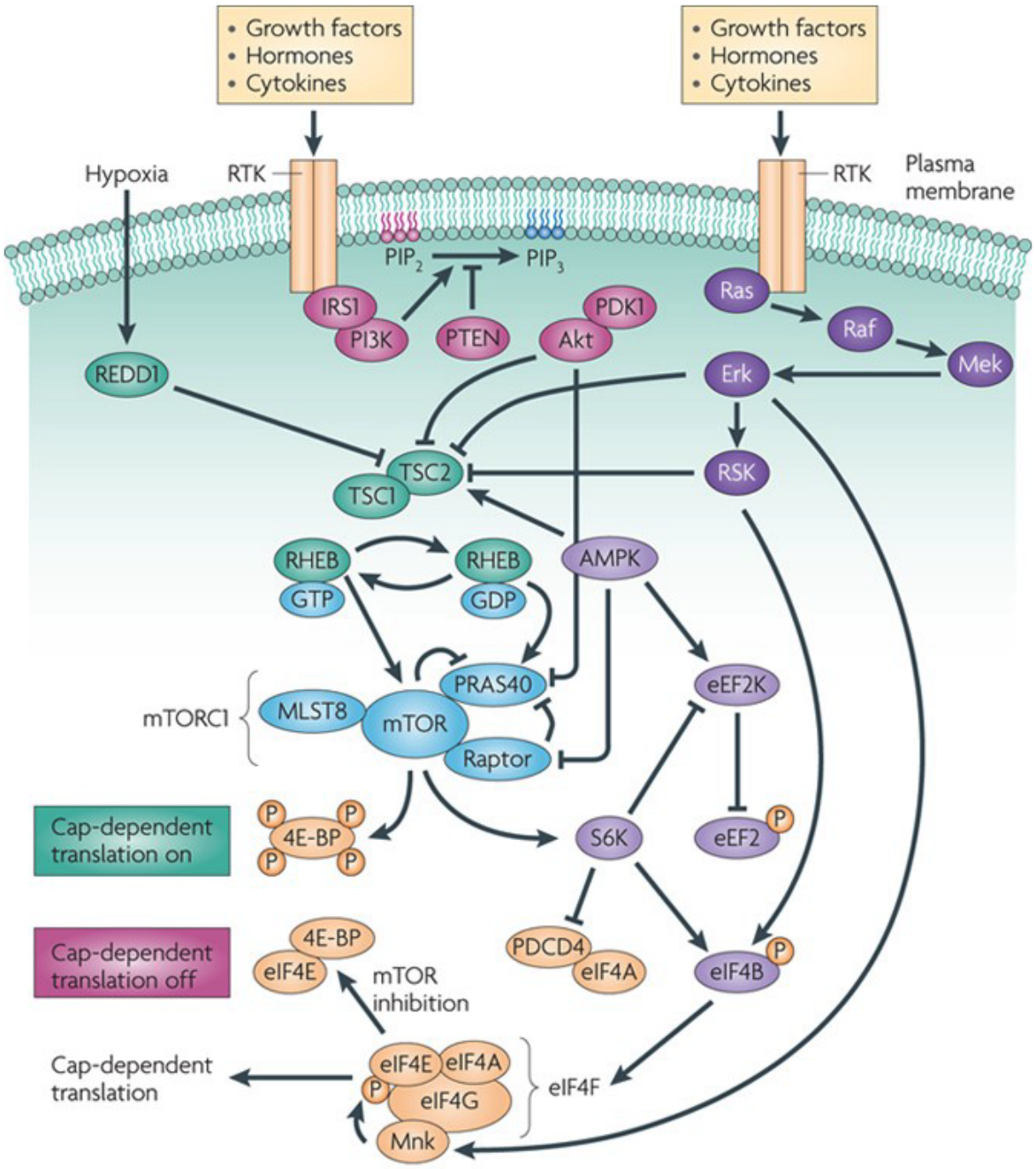

Figure 1.7: Signalling pathways converging at regulating protein synthesis
Mitogenic stimulation (growth factors, hormones or cytokines) targets receptor tyrosine kinases (RTKs), which promotes RAS/ERK and PI3K/PTEN/AKT signalling. ERK activates 5′cap dependent translation either through MNK-mediated phosphorylation of eIF4E, RSK-mediated phosphorylation of eIF4B or by inhibiting TSC1/TSC2complex, a negative regulator of mTOR signalling. AKT also regulates translation through mTOR by inhibiting another negative regulator, PRAS40. From Silvera et al. (2010)

Although the contribution of p53 to tumour formation is one of the best studied mechanisms of cancer, its role in controlling protein synthesis and being regulated by translation apparatus is less known. Overall, activation of p53 leads to suppression of ribosome biogenesis affecting translation initiation. Ribosome biogenesis is a complex process that coordinates the synthesis of rRNA, ribosomal proteins and other auxillary factors; p53 can control the activity of all of them. Firstly, p53 directly interferes with the assembly of RNA Pol I complex, which is essential for rRNA transcription (Zhai and Comai, 2000). Secondly, lessons learnt from studying ribosomopathies show that there is a bidirectional feedback loop between p53 and the expression of ribosomal proteins. When the process of ribosome biogenesis is impaired, for example due to mutation in a ribosomal protein, DNA damage of the ribosomal DNA (rDNA) or insufficient transcription of rRNA, free ribosomal proteins are released to the nucleoplasm. RPL5, RPL11, RPL23a, RPS7, and RPL26 are known to interact and sequester MDM2, a key inhibitor of p53 (Kampen et al., 2020) This leads to cell cycle arrest, induction of cellular senescence, or apoptosis, but it can vary between cell types. For example, haploinsufficiency of RPS14 or RPS19 in erythroid progenitor cells is known to increase p53 and CDKN1A (p21) protein *in vivo* to the level similar to those after gamma irradiation (Dutt et al., 2011). p53 can inhibit transcription of ribosomal proteins and other associated factors, indirectly, by inhibiting MYC (Ho et al., 2005). The dynamic of p53-related nucleolar stress response and its links to cancer and neurodegenerative diseases are not fully understood; this topic was reviewed recently by Lindström et al. (2018) and Pfister (2019).

1.3.2 Ribosome biogenesis and its oncogenic potential

The involvement of ribosome biogenesis in oncogenesis is well understood in the context of MYC expression deregulation, see **section 1.3.1**, but it is not limited to it. Other ribosome biogenesis regulators with known oncogenic potential include netrin 1 (NTN1) and epithelial cell-transforming sequence 2 oncogene (ECT2). ECT2 and truncated NTN1 isoform, expressed exclusively in cancer cells, promote rDNA transcription and pre-rRNA processing. rDNA is one of the most transcriptionally active regions of the genome and rDNA rearrangements are observed in the majority of lung and colorectal cancers (Stults et al., 2009). During S phase, when rDNA transcription continues, RNA polymerase I collides with the replication forks, which facilitates the formation of R-loops (rRNA-rDNA hybrids) (Pelletier et al., 2018). R-loops are known hot spots of DNA damage (Helmrich et al., 2011; Skourti-Stathaki and Proudfoot, 2014) and rDNA regions are common fragile sites (Pelletier et al., 2018).

Cancer-specific changes in the ribosome morphology are another interesting concept, which is sometimes referred to as the *oncoribosome* hypothesis. Varying expression level of ribosomal proteins, post-transcriptional rRNA modifications and recurrent mutations in ribosomal proteins are observed in many malignancies (Babaian et al., 2020; Bastide and David, 2018; Pelletier et al., 2018)

1.3.3 Translation factors are frequently deregulated in cancer

Regulation of translation initiation is essential for maintaing cell homeostasis during malignant transformation and under exposure to unfavorable conditions.

High levels of eIF4E are common in cancer and *in vitro* studies revealed that this can be sufficient to induce malignant transformation (Lazaris-Karatzas et al., 1990; Gingras et al., 1999). Oncogenic properties of eIF4E has been shown also *in vivo*. Transgenic mice overexpressing eIF4E have increased risk of several types of cancers, including B-cell lymphomas (Ruggero et al., 2004) and accelerated B-cell lymphomagenesis when *c-Myc* is deregulated as well (Wendel et al., 2004). Oncogenic properties of high levels of eIF4E are attributed to its ability to regulate translation of specific transcripts rather than control on global translation. Haploinsufficiency of eIF4E is associated with normal translation levels and development, but still prevents HRAS-induced transformation in mice (Truitt et al., 2015). eIF4E-sensitive mRNAs (eIF4E regulon) are enriched for CERT motif and complex RNA structures sequences in their 5′UTRs. The latter involves eIF4E-mediated recruitment of RNA helicase - eIF4A (Smith et al., 2021). These include transcripts associated with cell proliferation, survival and oxidative stress, such as cyclins, ornithine decarboxylase (ODC), vascular endothelial growth factor (VEGF), MYC, and phosphoribosyl-pyrophosphate synthetase 2 (PRPS2) (Bhat et al., 2015). Tumourgenesis following eIF4E overexpression develops relatively late, suggesting that other prooncogenic events might be required for full transformation (Ruggero et al., 2004).

Interestingly, the role of eIF4E is not limited to translation initation. A substantial proportion of it exists withing the nuclei, where eIF4E regulates export of selected mRNA to the cytoplasm through the nuclear pore. Known transcripts, whose export is promoted by eIF4E, include MYC, BCL6 and BCL2 (Culjkovic-Kraljacic et al., 2016).

Other translation factors important from the oncogenesis perspective are eIF3 subunits. Prooncogenic properties have been reported for eIF3A–eIF3D, eIF3G, eIF3H, eIF3M, eIF3E and eIF3F, but the exact role of individual subunits is not fully understood, see Smith et al. (2021). Similairly to eIF4E, deregulation of certain eIF3 subunits leads to preferential translation of specific prooncogenic regulon, which includes JUN, genes associated with epithelial–mesenchymal transition or inflammation (Smith et al., 2021; Lee et al., 2016; Desnoyers et al., 2015). Some functions are independent on the role of eIF3 complex in translation initation. For example, the role of eIF3A and eIF3B encompass

control of enzymes responsible for reversible N6-methyladenine RNA modification (Smith et al., 2021). eIF3-complex independent pool of eIF3H acts as a deubiquitylating enzyme stabilising YAP1, which is important for tumour progression and metastasis (Smith et al., 2021). An interesting dual role has been shown also for eIF3G, which is cleaved during apoptosis and translocated to the nucleus, where promotes caspase activation and DNA degradation (Smith et al., 2021).

When 5′cap-dependent translation is inhibited following cellular stress, eIF2A and eIF5B promote translation of selected transcripts through non-canonical, IRES-mediated initation of translation. These include few apoptosis inhibitors (XIAP, BIRC2 and BCL2L1), CDKN1A (p21) and proteins associated with NF-κB signalling (Smith et al., 2021).

The role of translation initation factors in cancer has been summarised in **Figure 1.8**

1.3.4 The significance of translational response to stress in cancer

Uncontrolled cell growth and proliferation combined with limited nutrients supply, hypoxic environment and deregulated cellular energetics are one of the hallmarks of cancer (Hanahan and Weinberg, 2011). Adaptation to unfavourable conditions is essential to initiate and maintain malignant transformation. Key to this process are evolutionarily conserved stress-signalling pathways that aim to rebalance cell homeostasis or commit it to apoptosis, if this cannot be achieved. Stress response encompasses a variety of mechanisms of which the Integrated Stress Response (ISR), Unfolded Protein Response (UPR) and heat shock response (HSR) are the ones most relevant for the topic of protein translation.

The principal point of ISR is the control of the avaliability of the ternary complex for translation, thus reprogramming the global translation landscape (Costa-Mattioli and Walter, 2020). The ternary complex consists of three subunits: eIF2α, eIF2β and eIF2γ. Accumulation of misfolded proteins, amino acid deprivation and other stress signals activate eIF2α kinases: PERK, PKR, HRI, and GCN2 that converge on the phosphorylation of eIF2α.

The role of the ternary complex is AUG start codon recognition that triggers GTP hydrolysis of eIF2 with the aid of eIF5. eIF2-GDP dissociates from the 40S complex and is recycled for another round of translation by eIF2B, which restores eIF2-bound GDP to GTP (Pakos-Zebrucka et al., 2016). The α subunit of eIF2 is the main point of control. In response to various stress stimuli, it can be phosphorylated at Ser 51 (P-eIF2α) and act as an inhibitor of eIF2B. In its inactive form (GDP-bound) P-eIF2α is limiting for the ternary complex formation, which results in the reduction of global, 5′cap dependent protein synthesis enabling 5′cap independent translation of selected transcripts.

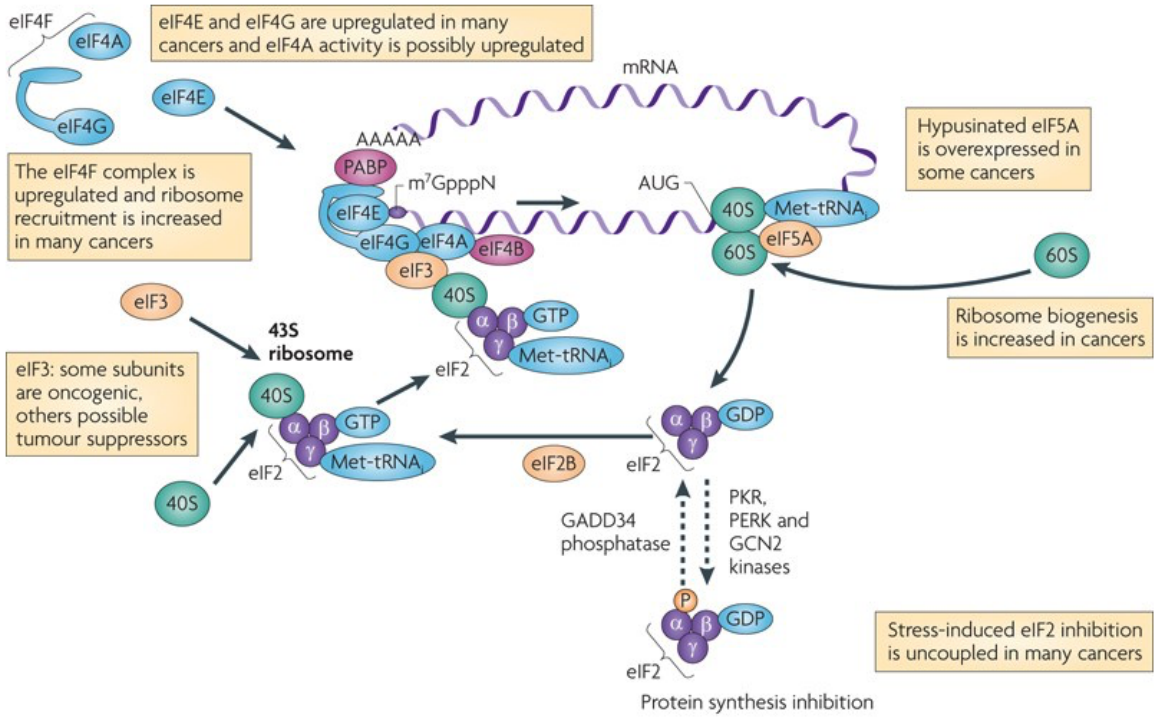

Figure 1.8: Overview of 5′cap dependent translation initiation in cancer
Numerous components of the translation initation machinery are deregulated in human cancers. Up to now, the link between oncogenesis and translation initiation has been well established for the components of eIF4F, eIF3 and eIF2 complex and ribosome biogenesis. The overexpression of **eIF4F** components, which include eIF4E, eIF4A and eIF4G results in preferential translation of selected transcripts. For individual **eIF3** subunits similar mechanisms was observed. **eIF2**α is phosphorylated by different kinases (PKR, PERK or GCN2), which are reactive to various types of stress. The net effect of P-eIF2α is repression of 5′cap-dependent translation, promotion of non-canonical translation initation and adaptation to unfavourable conditions. **Ribosomal biogenesis** is another mechanisms *hacked* by cancer cells to drive a malignant phenotype. Many oncogenic signalling pathways converges at controlling the synthesis of the ribosome components.
Other factors with possible role in tumourgenesis include eIF5A, eIF5B and eIF2A. From (Silvera et al., 2010)

Of these, activating transcription factor 4 (ATF4), activating transcription factor 5 (ATF5), DNA Damage Inducible Transcript 3 (DDIT3 or CHOP) promote transcription of genes responsible for restoration of cell homeostasis (Hinnebusch, 2005; Vattem and Wek, 2004; Pakos-Zebrucka et al., 2016).

The presence of misfolded proteins in the endoplasmic reticulum (ER) or the cytosol can have serious consequences for the cell function. The salient role of UPR is to sense unfolded proteins in the ER, while HSR respond from the cytosol increasing protein folding and degradation capacity (Costa-Mattioli and Walter, 2020). UPR consists of three main sensors: inositol-requiring protein 1 (IRE1α), protein kinase RNA-like endoplasmic reticulum (ER) kinase (PERK) and activating transcription factor 6 (ATF6), which controls the folding capacity of the ER (Hetz, 2012). Early UPR, mediated by PERK-eIF2 and converging with ISR, aims to reduce translation intensity decreasing protein load to the ER. This is accompanied by IRE1-dependent decay of mRNA and activation of the autophagy pathway (Hetz, 2012). Then, a group of transcription factors: ATF4, transcription factor 6 cytosolic fragment (ATF6f) and spliced X box-binding protein 1 (XBP1s) trigger transcription of adaptive genes aiming to alleviate stress and increase protein folding capacity. If the exposure and duration of stress exceeds the cell's ability to restore homeostasis, the cell commits to apoptosis (Hetz, 2012).

Stress response signalling controls pro-survival and apoptotic pathways, so the role of the stress response in cancer is complex and context specific **Figure 1.10**, reviewed by Clarke et al. (2014) and Urra et al. (2016). Briefly, many oncogenic pathways drive hyperactivation of protein synthesis making the cell susceptible to stress-induced apoptosis. However, high levels of UPR activation are associated with malignant transformation (Urra et al., 2016), are known to promote resistance to chemotherapy and radiotherapy in different cancer types (Rouschop et al., 2010; Ghaddar et al., 2021), correlate with poor prognosis in glioblastoma, breast cancer and pre-B acute lymphoblastic leukemia, promote metabolic reprogramming in prostate cancer and immunosuppressive environment in ovarian cancer (Clarke et al., 2014; Urra et al., 2016).

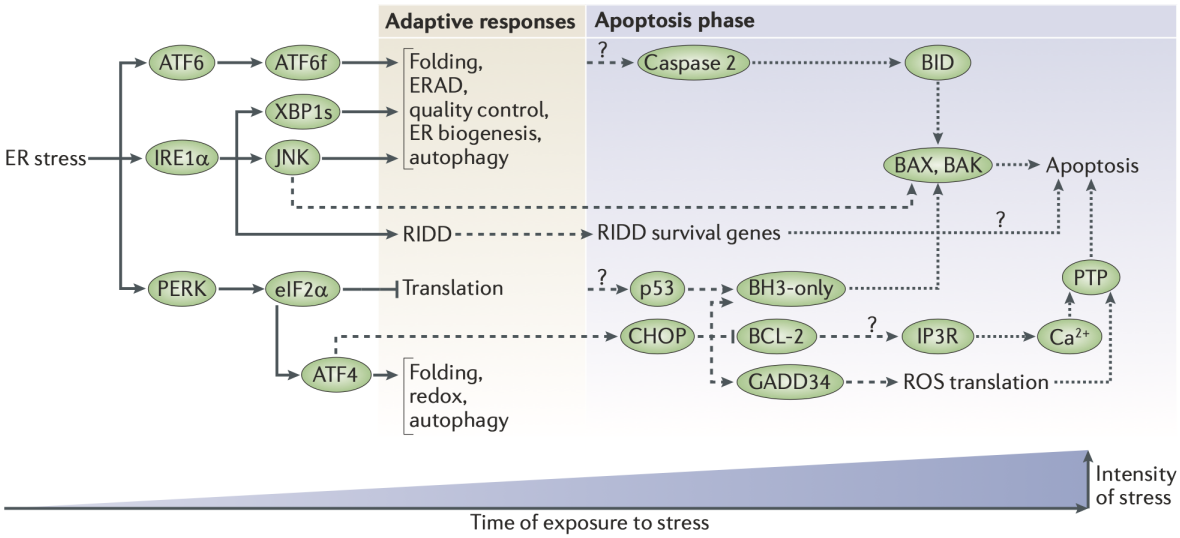

Figure 1.9: Overview of the cell fate decisions associated with ER stress response
The outcome of ER stress response involves adaptation and restoration of cell's protein folding capacity or triggering apoptosis, if the stress is prolonged or not restrained. There are three signalling arms responsible for the initiation of ER stress response: ATF6, IRE1α or PERK-mediated. Of which the latter, is one of the component of the Integrated Stress Response pathway. The net effect of the ER stress activation is transcriptional upregulation of several ER chaperones to facilitate protein folding. ER stress also include inhibition of 5′cap dependent translation (through eIF2α) and activation of ER-associated degradation (ERAD) pathway, autophagy to remove misfolded proteins and regulated IRE1-dependent decay (RIDD) pathway, which selectively removes mRNAs encoding proteins located in the endoplasmic reticulum (Hetz, 2012). Above certain threshold, the cell's ability to tolerate ER stress may be exceeded and ER stress response signalling will trigger the apoptosis. In such scenario, cell death is controlled by BCL2 protein family, BAX and BID and p53-mediated signalling.
From Hetz (2012)

1.4 Elements of B-cell biology from the perspective of lymphoma development

The current World Health Organisation (WHO) classification of B-cell malignancies relies predominantly on the cell of origin of the tumour. B-cell neoplasms can be divided into malignancies originating from either progenitor cells, mature B-cells or plasma cells (Swerdlow et al., 2016). Overall, the majority of Non-Hodgkin Lymphomas (NHL), which are the most common haematological malignancies in adults, are derived from the germinal center (GC) stage of B cell development (Mlynarczyk et al., 2019; Swerdlow et al., 2016). GC-derived lymphomas include: Diffuse Large B-cell Lymphoma (DLBCL), Burkitt Lymphoma (BL) and Follicular Lymphoma (FL), which all account for about 60% of all B-cell NHL.

GCs are transient microstructures, which are formed in secondary lymphoid organs upon activation of mature naive B-cells by T-cell dependent antigens. The GC reaction is dedicated to selecting and expanding populations of B-cells that will eventually differentiate into memory B-cells, and plasma cells. This is an essential part of a physiological adaptive immune response against exogenous pathogens (Klein and Dalla-Favera, 2008; De Silva and Klein, 2015). Lack of GCs observed in patients with inherited hyper-IgM syndrome is associated with severe immunodeficiency, which highlights the importance of the GC reaction (Etzioni and Ochs, 2004).

The GC consists of two functionally and histologically distinct zones: the dark zone and the light zone. The dark zone consists of a dense population of large B-cells - centroblasts, that undergo rapid proliferation and somatic hypermutation (SHM). The aim of the SHM process is to introduce point mutations into the variable regions of the heavy and light chains of the immunoglobulin genes (IgV), that will generate a population of mutant subclones with a broad range of B-cell receptor (BCR) affinity for the antigen. This is achieved through expression of activation-induced cytidine deaminase (AICDA) (Muramatsu et al., 2000; De Silva and Klein, 2015). Centroblasts then migrate to the light zone becoming centrocytes. Centrocytes compete for survival signal from CD4 T follicular helper (T_{FH}) cells and Follicular Dendritic Cells (FDC) and only those demonstrating high-affinity towards the immunising antigen evade apoptosis (Nakagawa and Calado, 2021). The positive selection (survival) stimuli is provided by B-cell receptor (BCR) capturing the FDC-bound intact antigen, co-stimulatory receptors such as CD40 or tumour-necrosis factor (TNF)-receptor, adhesive molecules (ICAM-1 and VCAM-1), anti-apoptotic molecules (BAFF), and a mixture of cytokines secreted by FDC and T_{FH}. BCR-antigen complex is internalised and the antigen is then presented at the cell surface through MHC II complex that interacts with CD4 T follicular helper cells. The higher the affinity the more antigen can be acquired from FDC and the stronger the stimulation received from T_{FH}

is (Nakagawa and Calado, 2021). Immunoglobulin Class Switch Recombination (CSR) is another important event in the life of GC B-cells. The mechanism of CSR involves double stranded breakage and rejoining of the immunoglobulin genes leading to deletional recombination of the heavy chain region, thus determining the final class of produced antibodies (IgG, IgE or IgA). GC B-cells undergo several rounds of proliferation and selection, re-cycling between dark and light zone through a process directed by chemokines gradients. Selected GC B-cells differentiate into first antibody-secreting plasma cells after about one week since activation (Basso and Dalla-Favera, 2015; Shaffer III et al., 2012). The affinity of the antibodies increases with time in the phenomenon known as "affinity maturation", which continues over the next weeks. An overview of the most important steps of B-cell development and maturation were shown in the **Figure 1.10**.

The family of GC-derived lymphomas is highly heterogeneous comprising a broad spectrum of clinical presentation and complex molecular context. Each lymphoma subtype works like a distorting mirror, which resembles a certain stage of B-cell development, sometimes referred to as *cell-of-origin*, but aggravates or attenuates specific physiological mechanisms to drive a malignant phenotype. The oncogenic events may start at earlier stages than the normal B-cell counterpart and accumulate during differentiation and maturation (Shaffer III et al., 2012). For example, although FL resembles GC stage of development but initial genetic corruption begins at bone marrow pro-B cell stage (Shaffer III et al., 2012). In addition to genetic aberrations commonly found in tumours, such as point mutations, deletions or amplifications that activate potential oncogenes or deactivate tumour suppressor genes, there are two GC-specific processes, SHM and CSR, that largely contribute to the uniqueness of the lymphoma genome (Basso and Dalla-Favera, 2015; Klein and Dalla-Favera, 2008; Mlynarczyk et al., 2019). SHM and CSR occur in a highly hazardous environment, in which DNA damage response and cell cycle checkpoints are desensitised allowing for remodelling of the immunoglobulin loci and the expansion of selected clones. An unwanted by-product of CSR is the risk of translocation, which is common in GC-derived lymphomas. Chromosomal abnormalities in GC-lymphomas usually involve translocation of a proto-oncogene into an immunoglobulin locus which is under control of an active enhancer. This drives high-level and sustained transcription of the translocated oncogene (Küppers and Dalla-Favera, 2001). Thus translocation of *MYC* into the immunoglobulin loci, t(8,14), is a hallmark of Burkitt Lymphoma (occurs in more than 95% of cases) (Basso and Dalla-Favera, 2015; Klein and Dalla-Favera, 2008; Mlynarczyk et al., 2019). Other genes with recurrent translocation include *BCL2* (80% of FL patients) and *BCL6* (30–40% of DLBCL).

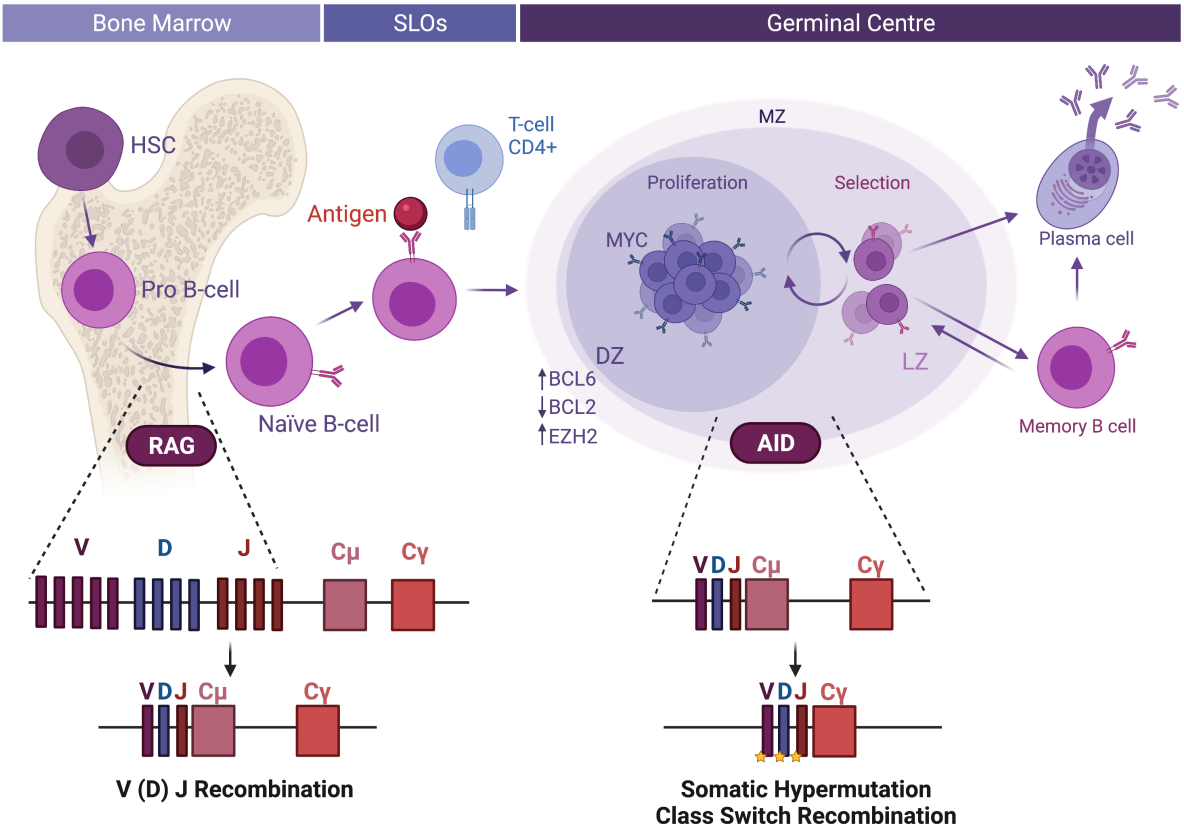

Figure 1.10: Overview of B-cell development and maturation

B-cells develop in bone marrow from haematopoietic stem cells (HSC). Recombination of the immunoglobulin loci is the first step in the generation of a mature immunoglobulin receptor. This process starts in the pre-pro-B cell and continues to the pro-B cell stage, leading to a heavy chain protein formation. V(D)J segments recombination requires the activity of the recombination-activating gene (RAG) complex, which directs cleavage of the DNA. Mature B-cells (naïve B-cells) leave the bone marrow and migrate to secondary lymphoid organs (SLOs), where they encounter an antigen. Once activated, with T-cell dependent antigen, B-cells enter the germinal centre reaction (De Silva and Klein, 2015). In the dark zone (DZ), activated B-cells, now centroblast, undergo proliferation and somatic hypermutation of their immunoglobulin loci. Centroblasts migrate to the light zone (LZ), becoming centrocytes and compete for the survival signal from CD4 T follicular helper (T_{FH}) cells and Follicular Dendritic Cells (FDC). During germinal centre (GC) GC reaction B-cells switch their isotype during class switch recombination. Editing of the immunoglobulin loci during somatic hypermutation and class switch recombination is mediated by activation-induced cytidine deaminase (AID) activity (De Silva and Klein, 2015). GC B-cells undergo several rounds of proliferation and selection, re-cycling between LZ and DZ. Finally, GC B-cells differentiate into memory B-cells or antibody secreting plasma cells (Mlynarczyk et al., 2019; Runte et al., 2019).

A massive phenotype shift from quiescent näive B-cells cells to rapidly proliferating and hypermutated GC B-cells and, finally, to antibody producing plasma cells is directed by profound reorganisation of the gene expression programme. Transcription factors with well established links to GC formation and lymphomagenesis include MYC and BCL6.

In the physiological GC reaction, the proto-oncogene MYC has a bimodal pattern of expression: it is induced early during GC reaction initation and then transiently in the LZ B-cells undergoing positive selection and DZ re-entry (Calado et al., 2012; Dominguez-Sola et al., 2012; Basso and Dalla-Favera, 2015). Expression of MYC primes B-cells for massive proliferation and clonal expansion. BCL6, known as a master regulator of GC-reaction, is key to GC formation and maintenance (Basso and Dalla-Favera, 2015). It acts as a transcriptional repressor inhibiting DNA damage response, apoptosis and premature B-cell activation and terminal differentiation into plasma cells (Basso and Dalla-Favera, 2015). BCL6 controls the expression of several genes known to drive tumour formation, such as cell cycle checkpoints (CDKN1A, CDKN1B), DNA damage response regulators (ATR, TP53) and differentiation factors (eg. IRF4 and PRDM1) (Ci et al., 2009; Basso and Dalla-Favera, 2015). Desensitised DNA damage response by BCL6 is necessary to sustain immunoglobulin loci remodelling and affinity maturation. The expression of BCL6 is promoted during the initation of GC reaction by IRF8 and MEF2B, OCT2 and OCA-B and then downregulated at the GC exit stage (Song et al., 2021).

Despite tremendous progress that has been made in our understanding of processes governing GC reaction and their links to lymphomagenesis, translation of these findings into concise subtype classification systems and then further into clinical practice has been challenging. Of all GC-derived lymphomas, DLBCL has been the most difficult to subdivide due to its considerable molecular heterogeneity. After years of attempts to classify DLBCL morphologicaly with little success, the first modern classification utilising the technology of next generation sequencing was the Cell Of Origin system (Alizadeh et al., 2000), which distinguished two DLBCL subtypes based on gene expression profile. Germinal centre B-like DLBCL (GCB-DLBCL) group resembling normal germinal centre B-cells and activated B-like DLBCL (ABC-DLBCL) similar to activation of peripheral blood B cells.

In addition to gene expression profile, certain genetic hallmarks of B-cell lymphoma were also recognised. Identification of *BCL2*, *MYC* and/or *BCL6* chromosomal rearrangements has led to distinguishing double/triple hit high grade B-cell lymphoma. This finding was further complemented by discovery of distinct transcriptomic signatures of MYC-driven GCB-DLBCL, that overlapped with double hit lymphoma cases, referred to as Molecular High-Grade DLBCL or DHSig, respectively (Sha et al., 2019; Ennishi et al., 2019).

More recently three genomic studies from Harvard (Chapuy et al., 2018), National

Cancer Institute (NCI) (Schmitz et al., 2018) and the UK Haematological Malignancy Research Network (HMRN) (Lacy et al., 2020) identified converging genetic subclasses based on the profile of mutations. The current status molecular profiling of DLBCL has been summarised in the **Figure 1.11** and covered in detail by Cutmore, Krupka, Hodson (2022)

1.5 Role of translation in B-cell development and malignancy

Although GC reaction and lymphomagenesis have been extensively studied at the level of transcription, much less is known about post-transcriptional regulation of B-cells development and maturation. Deregulation of mRNA translation is common in B-cell lymphoma and typically involves 1) changes in the expression of the core components of translation machinery or pathways regulating its activity, 2) microRNAs-driven regulation of protein synthesis, and 3) translational control of key lymphoma oncogenes or tumour suppressors (Horvilleur et al., 2010).

Aberrant MYC or mTOR signalling are the most studied examples of pathways affecting the proteins synthesis in B-cell lymphomas. MYC translocation and mutations are hallmarks of Burkitt Lymphoma and Molecular High-Grade (and Double-Hit) DLBCL subtype (Schmitz et al., 2012; Sha et al., 2019), while mTOR activation is usually associated with activation of B-cell receptor (BCR) signalling, PTEN loss (about 14 % of DLBCL), or mutations in ERK/RAS pathway. (Chapuy et al., 2018; Reddy et al., 2017; Lacy et al., 2020; Pfeifer et al., 2013)

The oncogenic activity of MYC includes hyperactivation of protein synthesis, but it is also dependent of the translation machinery. Haploinsufficeincy in *Rpl24* was shown to abrogate MYC-induced transformation *in vivo*, while maintaining normal B-cell development (Barna et al., 2008). This effect was specific to MYC-driven tumours as in $Tp53^{-/-}$ mice, the frequency and latency of tumours was not affected by $Rpl24^{+/-}$ (Barna et al., 2008). Simultaneous deregulation of c-MYC and eIF4E expression accelerates the development of B-cell lymphoma in transgenic mice. The molecular mechanism of this cooperation results from eIF4E's ability to suppress MYC-induced apoptosis by promoting cellular senescence (Ruggero et al., 2004). It is interesting in the light of another study, where overexpression of eIF4E in lymphoma was associated with chemotherapy resistance (Wendel et al., 2004).

Another important pathway converging at regulating protein synthesis is BCR signalling, which is vital to the survival of normal and malignant B-cells. Depending on whether it requires antigen engagement or not, BCR signalling in lymphoma is termed chronic active or tonic.

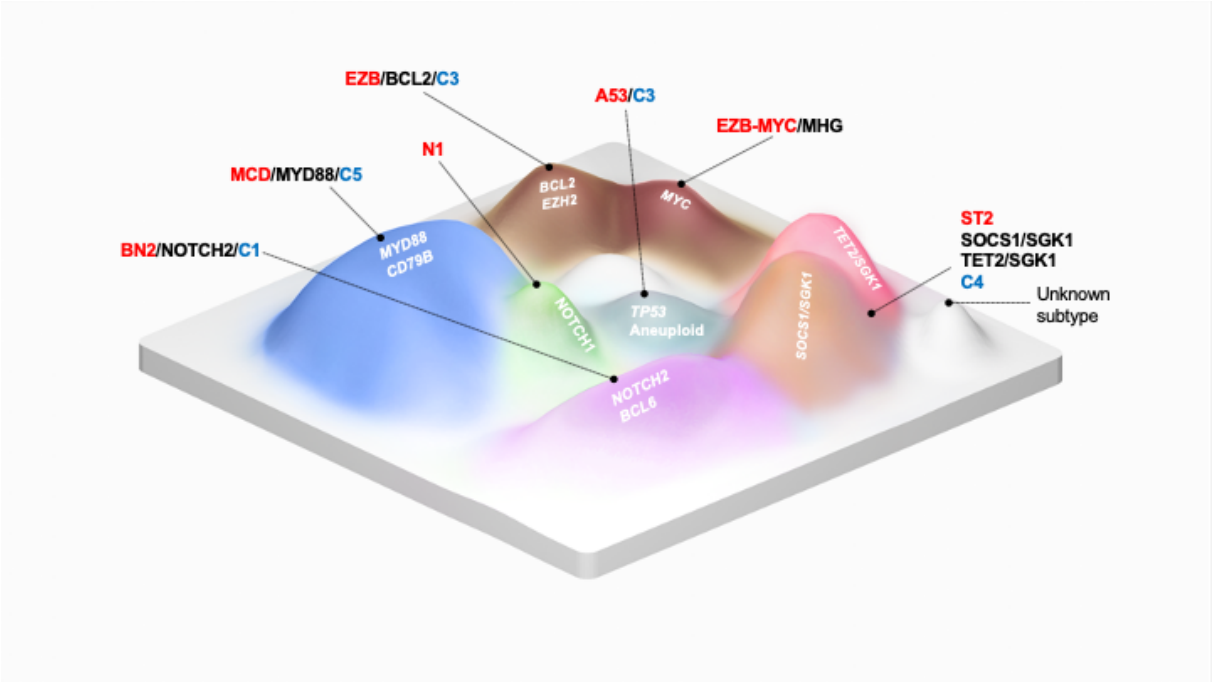

Figure 1.11: Topography of DLBCL molecular subtypes

A conceptual representation of the genetic classification of DLBCL. Coloured hills depict known genetic subtypes. The genes most commonly altered are indicated in white. Each subtype is labelled with LymphGen/NCI cluster name (red) (Schmitz et al., 2018), HMRN (black) (Lacy et al., 2020) and Harvard (blue) equivalent (Chapuy et al., 2018). Briefly, the **MCD/C5/MYD88** subtype covers DLBCL cases with poor prognosis driven by mutations in B-cell receptor, Toll like receptor (TLR) and NFκB pathway accompanied by immune evading mutations. The **EZB/C3/BCL2** group distinguishes cases with recurrent *BCL2*, *EZH2*, *CREBBP*, *KMT2D* mutations sharing the pattern with follicular lymphoma (FL). Interestingly, based on the *gene expression* signature or *MYC* amplification status, LymphGen and HMRN characterise an extra subtype of termed **EZB-MYC/MHG**. Next, the **BN2/C1/NOTCH2** subtype includes mutation of *NOTCH2*, *TNFAIP3*, *BCL10*, *NFKBIZ*. The **ST2/C4** group is characterised by activation of the JAK/STAT/ERK pathway and mutations in *SOCS1*, *DUSP2*, *STAT3* and *BRAF*. The HMRN systems additionally distinguishes two subgroups within ST2/C4, based on *SGK1* and *SOCS1* mutation status. Lastly, the N1/NOTCH1 subtype is defined by activating mutations of *NOTCH1* gene and the **A53/C3** subtype is characterised by enrichment for *TP53* mutations and aneuploidy. Whilst patients position on top of the coloured *hills* will be reproducibly classified by each classification system, patients in valleys may be unclassified or classified alternatively across different classification systems. Uncoloured hills correspond to unknown DLBCL subtypes which may emerge in the future from currently unclassified cases. See Cutmore, Krupka, Hodson (2022)

Whereas the pro-survival effect of the antigen-independent tonic BCR signalling is dependent on the activated PI3K/AKT pathway, the antigen stimulated chronic active BCR signalling engages multiple pathways including PI3K and NF-κB (Shaffer III et al., 2012). Tonic BCR activity is characteristic to BL and GCB subtype of DLBCL, while the chronic active BCR signalling is characteristic to ABC-DLBCL. BCR-mediated mTOR activation promotes both: global 5′cap dependent translation initation and preferential translation of selected transcripts through increased translation factors activity, see section 1.2.1 and 1.3.1 .

Although, core components of the translation machinery are rarely mutated in B-cell lymphoma (usually in less than 10% of cases), their expression is frequently abnormal (Taylor et al., 2020). The analysis of mRNA expression of 16 eIFs showed that 12 out of 16 are overexpressed in DLBCL compared to normal tissue samples (Unterluggauer et al., 2018). Deregulated activity of translation factors may drive pro-oncogenic programmes of translation. For example, overexpression of eIF4B promotes translation of proteins associated with tumour cell survival (DAXX, BCL2 and ERCC5) and correlates with poor clinical outcome (Horvilleur et al., 2014). Another study has shown that translation of key lymphoma oncogenes, such as MYC, BCL6, BCL2 is under the control of USP11, which is recruited to the translation initiation complex where it deubiquitinates and stabilises eIF4B (Kapadia et al., 2018).

In addition to the core components of the translation machinery, there are other RNA-binding proteins that, when mutated or deregulated, may promote lymphoma development and progression. Of these a DEAD-box helicase 3, X-linked (DDX3X) has attracted much attention recently as it been found recurrently mutated in Burkitt Lymphoma (Grande et al., 2019; Schmitz et al., 2012), Chronic Lymphocytic Leukaemia (CLL) (Ojha et al., 2015; Takahashi et al., 2018), medulloblastoma (Jones et al., 2012; Pugh et al., 2012; Robinson et al., 2012), head and neck squamous cell carcinoma (Stransky et al., 2011) and NK-T cell lymphoma (Jiang et al., 2015). The role of DDX3X has been extensively studied in medulloblastoma (Jones et al., 2012; Pugh et al., 2012; Robinson et al., 2012; Patmore et al., 2020; Samir et al., 2019), where it has been classified as tumour suppressor gene regulating the activity of Wnt signalling. DDX3X function in other cancers, including lymphoma, is less understood as it has been recognised as both tumour suppressor or oncogene depending on the cancer type, what is not surprising given ubiquitous role of DDX3X in RNA biology. DDX3X roles include regulation of transcription, splicing, nuclear export, stress granule formation and resolution, microRNA biogenesis, mRNA translation and decay (Linder and Jankowsky, 2011; Mo et al., 2021). It is entirely unknown which of these functions may be relevant to lymphoma.

Table 1.1: Examples of translationally regulated genes relevant to mature B-cell lymphomas, from Taylor et al. (2020)

Gene(s)	Protein function	Association with B-cell neoplasm	Evidence for translational control	**Ref**
MYC	Cell growth	t(14,8) translocation in BL Recurrent mutations in MHG/Double-Hit DLBCL Increased mRNA translation following BCR stimulation	Inhibition of eIF4A reduces MYC expression in lymphoma cell lines Inhibition of eIF4E reduces MYC mRNA translation and nuclear export in DLBCL cell lines	(Yeomans et al., 2016) (Culjkovic-Kraljacic et al., 2016) (Steinhardt et al., 2014) (Schatz et al., 2011) (Wilmore et al., 2021)
MCL1	Cell survival (anti-apoptotic)	Expression induced following BCR stimulation	Inhibition of eIF4A reduces MCL1 expression in lymphoma cell lines	(Schatz et al., 2011)
BCL2	Cell survival (anti-apoptotic)	t(14;18) translocation in FL Recurrent translocations in GCB-DLBCL and double/triple hit DLBCL	Inhibition of eIF4E reduces BCL2 mRNA translation and nuclear export in DLBCL cell lines	(Culjkovic-Kraljacic et al., 2016)
CCND1	Cell cycle progression	t(11;14) translocation in mantle cell lymphoma	Inhibition of eIF4A reduces CCND1 expression in lymphoma cell lines	(Schatz et al., 2011)
BCL6	Transcriptional repression	t(14;18) translocation in double/triple hit lymphoma	Inhibition of eIF4E reduces BCL6 mRNA translation and nuclear export in DLBCL cell lines	(Culjkovic-Kraljacic et al., 2016)
CARD11 ***BCL10*** ***MALT1***	CBM complex components which control NF-κB activation	Recurrent mutations in subset of ABC-DLBCL Mediates NF-κB activation downstream of BCR	Inhibition of eIF4A reduces CBM complex expression in lymphoma cell lines	(Steinhardt et al., 2014)

Another interesting topic is aberrant expression of microRNA (miRNA) in B-cell malignancies. MiRNAs are small (about 22 nucleotides) non-coding RNAs that interact with a group of mRNA, through fully or partially complementary sequence in the 3′UTR, regulating their translation activity and degradation, see Filipowicz et al. (2008) for a review. Expression of specific miRNAs signatures in DLBCL predict lymphoma subtype, event-free survival and the risk of transformation from indolent follicular lymphoma to DLBCL (Lawrie et al., 2009; Malumbres et al., 2009; Li et al., 2009a). Characteristic miRNAs patterns have been also observed for Burkitt Lymphoma (Leucci et al., 2008), Hodgkin Lymphoma (Navarro et al., 2008) and Mantle Cell Lymphoma (Zhao et al., 2010). Interestingly, even for the same disease, little overlap is observed in miRNAs expression pattern between individual studies. miRNAs identified as important for one type of lymphoma do not necessarily have similar effect in another, or their role might be ambiguous (Horvilleur et al., 2010). The role of individual miRNA in B-cell malignancies has been reviewed by Sole et al. (2016).

Finally, the expression level of several genes relevant to B-cell malignancies can be regulated at the level of translation, detailed in **Table 1.1**.

1.6 Toolkit to study heterogeneity of translation

"A small particulate component of the cytoplasm" was the title of the paper, where George E. Palade revealed the ribosome to the world (Palade, 1955). Although the process of translation has been known since 1955 and its biochemical and molecular nature was comprehensively studied over the last decades, the analysis of the dynamics of these "small particles", especially from a genome-wide perspective, is still challenging. Technical difficulties associated with studying translation are related to the necessity of studying simultaneously two chemically distinct types of molecules: RNA and protein. The translation process refers to the very short and dynamic moment when the two families of molecules physically interact and one (RNA) catalyses the synthesis of another. Therefore measuring the concentration of total transcript or protein abundance does not reflect *true* translation intensity as many factors, not directly related to protein synthesis, may affect the abundance of both. A simple workaround is to use special separation or labelling techniques to extract subpopulation of mRNA or protein that are functionally related to translation. At the protein level, it will be a pool of newly synthesised proteins, ideally nascent proteins that are still bound by the ribosome. At the transcript level, the evidence of physical interaction with the translation machinery can be used as a proxy for active translation.

Another challenge is translation heterogeneity, which involves not only differences in translation intensity between single cells and tissues, but also varying translation

programmes of single or groups of transcripts. This phenomenon addresses the problem of the scale and resolution of translation quantification.

Available methods can be classified by the number of transcripts or proteins measured simultaneously into low-throughput and high-throughput, by the level of insight, e.g. bulk, single cell or cell compartments or by resolution - single nucleotide, transcript wide or global. The choice of the experimental technique is dictated by the biological question and the level of throughput required to answer it.

Net-abundance of each protein is a function of both - synthesis and degradation rate, so that it is safe to assume that only the pool of newly synthesised proteins can mirror the translation rate (Iwasaki and Ingolia, 2017). Because of relatively large content of pre-existing proteins in the cell, special separation or labelling techniques are needed to quantify this subpopulation.

Earlier methods involved radioisotope labelling of newly synthesised proteins, for example with ^{35}S-methionine or cysteine. Nowadays, non-radioactive, luminescent labelling is more popular. The most common techniques utilise aminoacyl tRNA analogues, such as azidohomoalanine (AHA) or alkyne-bearing homopropargylglycine (HPG) (Iwasaki and Ingolia, 2017), which are incorporated into the C-terminal end of nascent proteins. The disadvantage of using amino-acids analogues is methionine-depletion step, which can induce stress response and disrupt a delicate translation pattern. This does not apply to puromycin analogue - alkyne-bearing puromycin (O-propargyl-puromycin; OPP), which induces premature translation termination and release of nascent proteins with OPP attached. Labelled proteins, depending on the exact assay, can be detected by specific antibodies or secondary labelled with fluorophore by CLICK chemistry (Aviner, 2020). The protocol can be conjugated with several well-established techniques, such as confocal microscopy, fluorescence-activated cell sorting (FACS) or classical immunohistochemistry (Liu et al., 2012; Iwasaki and Ingolia, 2017). Methods involving nascent proteins labeling allows the measurement of bulk protein translation, or, depending on the exact protocol, provides single-cell or event cell compartment resolution.

High-throughput techniques have been dominated by two approaches: mass-spectrometry (MS) or mRNA sequencing based. The first usually combines some type of labelling to isolate newly synthesised proteins with subsequent measurement with tandem MS. Pulsed stable isotope labeling by amino acid in cell culture (pSILAC), which is the most popular technique, involves supplementation of the growing media with amino-acids labelled with stable isotopes only for a brief moment. Proteins with incorporated *heavy* amino-acids (newly synthesised) can be easily distinguished from pre-existing *light* proteins by their mass. The abundance of *heavy* labelled proteins reflect the relative intensity of protein

synthesis (Iwasaki and Ingolia, 2017). Other MS techniques, such as BONCAT, QuaNCAT or PUNCH-P, relies on isolation of labelled proteins with streptavidin beads. In bio-orthogonal noncanonical amino acid tagging (BONCAT), AHA-labelled proteins are tagged with biotin. A variation of BONCAT, quantitative noncanonical amino acid tagging (QuaNCAT), is a combination of BONCAT and SILAC allowing to quantify early changes of protein synthesis with higher specificity and greater depth than standard BONCAT (Howden et al., 2013). Finally, in puromycin-associated nascent chain proteomics (PUNCH-P) (Aviner et al., 2013), translating ribosomes are isolated from the cell with ultracentrifugation, next, nascent peptides are labelled with biotin-dC-puromycin, pulled down with streptavidin beads and analysed by MS (Aviner, 2020). An advantage of MS-based approaches is their accuracy and high-throughtput. Unfortunately, quantification of the protein abundance with MS has limited dynamic range and capacity to identify novel products, isoforms or proteins with post-translational modifications (Iwasaki and Ingolia, 2017).

Polysome fractionation and ribosome footprinting (a.k.a Ribosome Profiling or Ribo-Seq) are two main high-throughput strategies to analyse translation from the individual transcript perspective **(Figure 1.12)**. Both are almost equally popular nowadays, but their technical protocols differ substantially and thereby biological interpretation of the results. Ribo-Seq and polysome fractionation work under the assumption that 1) actively translated mRNAs are physically associated with the ribosomes, and 2) the efficiency of translation is proportional to the number of ribosomes bound to mRNAs (Piccirillo et al., 2014). In polysome fractionation, cell lysates containing polysomes (mRNAs with multiple ribosomes), monosomes (mRNAs with a single 80S ribosome), and free ribosomal subunits are centrifuged over a sucrose density gradient, which allows for separation of transcripts according to the number of ribosomes attached. This is referred to as a polysome profiling step. Each fraction can be distinguished by change in optical density (absorbance) and subjected to analysis. The composition of it can be evaluated in terms of the individual mRNA or protein abundance. For mRNA-centric studies, cDNA microarray (historical approach), RNA-seq (current approach) or RT-PCR (low-throughput) can be used (Piccirillo et al., 2014; Johannes et al., 1999; Karginov and Hannon, 2013). Typically, mRNAs translated more efficiently will be enriched in the heavy polysome fraction (more than three ribosomes per mRNA) (Piccirillo et al., 2014; Johannes et al., 1999), while less translated mRNA will be more abundant in the light fraction, which usually corresponds to polysomes with three or less ribosomes and the monosome fraction. For better resolution, each fraction can be quantified separately.

Ribo-Seq is centred on the idea of *ribosome footprinting*, developed initially by Wolin and Walter (1988), which allows to determine the position of a ribosome with single-nucleotide precision (Ingolia et al., 2009). Ribosomes bound to a mRNA protect a

fragment (about 30 nucleotides) from RNAse digestion (ribosome-protected fragments, RPFs). By utilising the advances in next-generation sequencing, it is possible to sequence those RPFs, thus, to infer the position of each translating ribosome. The information about the number and the location of ribosomal footprints can be used to compute ribosome density for a single transcript, which will be a proxy for its translation intensity (Ingolia et al., 2009). In contrast to Ribo-Seq, polysome profiling-based approaches do not allow to analyse the precise location of attached ribosomes. It might be enough to judge which transcripts forms polysomes (are actively translated), but no information about the exact location and reading frame are available. On the other hand, polysome fractionation maintains the entire transcript intact, which may be advantageous for studying the effect of differential splicing or polyadenylation on translation. Polysome fractionation combined with RNA-Seq reports the total number of ribosomes per mRNA, which depending on the fraction in which a transcript was found, while Ribo-Seq returns the position of every ribosome (regardless of the polysome/monosome fraction of the entire transcript). Hence, Ribo-Seq allows for quantification of the relative ribosome density per transcript, whereas polysome fractionation-RNA-Seq gives an estimation of the absolute number of ribosomes. This may prove less efficient when studying translation landscape following large shifts in global translation, but no systematic comparison has been performed so far (Piccirillo et al., 2014) Finally, with Ribo-Seq it is possible to annotate *de novo*, per sample, directly from the data, which regions of the genome are actively translated at subcodon level, see chapter 5, which is not possible with polysome profiling.

In a typical Ribo-Seq experiment, treatment with a translation inhibitor briefly before harvesting aims to *freeze* the translating ribosomes preventing them from dissociating from the transcript and maintaining the fidelity of the ribosome footprint localisation. The combination of ribosomal footprinting with different translation inhibitors, targeting different stages of translation, can broaden the biological application of the Ribo-Seq data. Cycloheximide (CHX), used in the original Ribo-Seq protocol (Ingolia et al., 2009; McGlincy and Ingolia, 2017), is the most popular choice. CHX is a small molecule inhibitor blocking eukaryotic translation elongation by binding to the ribosome E-site and blocking eEF2-mediated translocation (Obrig et al., 1971; Schneider-Poetsch et al., 2010; de Loubresse et al., 2014). In contrast, two other popular inhibitors: lactimidomycin (LTM) and harringtonine (HARR), require empty E-site, thus inhibit specifically only the first round of elongation capturing the ribosomes at the translation initiation site (Iwasaki and Ingolia, 2017; de Loubresse et al., 2014). Ribo-Seq combined with HARR or LTM treatment allows to build a genome-wide map of start codons and is known as global translation initation site sequencing (GTI-Seq) (Iwasaki and Ingolia, 2017). The ribosomes not blocked by the translation inhibitor dissociates from the transcripts leaving only the initiation site footprints for sequencing. A variation of this protocol, quantitative

translation initiation sequencing (QTI-seq), involves sequential treatment with LTM and puromycin, which facilitates run-off of non-initiating ribosomes increasing resolution of translation initiation site localisation (Gao et al., 2015). The full spectrum of ribosome footprinting-based techniques has been reviewed by Iwasaki and Ingolia (2017)

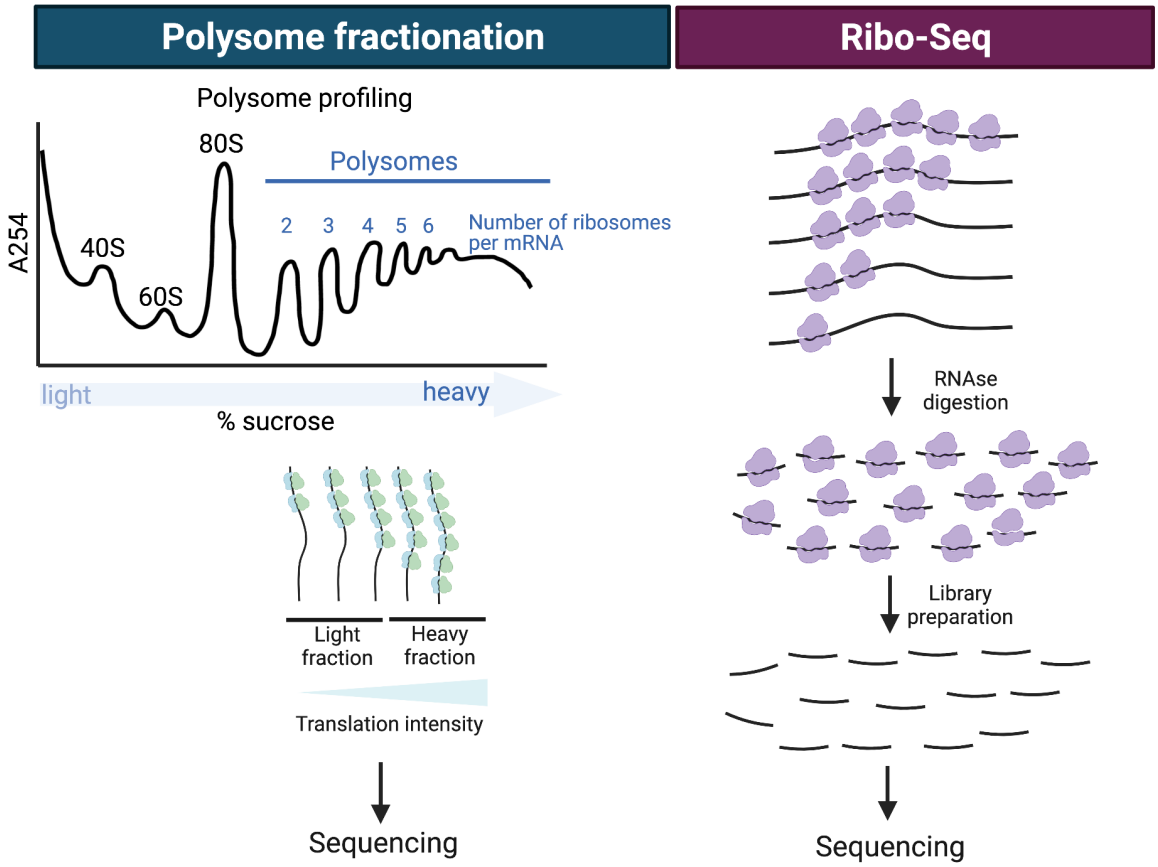

Figure 1.12: Comparison of polysome fractionation and Ribo-Seq technique
Polysome fractionation and Ribo-Seq allow analysing the translation intensity of individual transcripts. Polysome fractionation is based on the technique of polysome profiling, where a sucrose density gradient separates polysomes, monosomes and individual ribosomal subunits. Each polysome fraction contains the transcripts with a different number of ribosomes attached. By employing a next generation sequencing technology (e.g. RNA-Seq), it is possible to infer which transcripts are abundant in each fraction. In contrast, in the Ribo-Seq workflow, transcripts are digested with RNAse generating a mixture of mRNA fragments protected from RNAse digestion by attached ribosomes. These, so-called ribosomal footprints can be analysed with next generation sequencing technique.

1.7 Project aims

The aim of this thesis is to explore the role of translational control during lymphoma development.

Firstly, I established an automated bioinformatic pipeline for efficient processing of Ribo-Seq datasets which allows to keep the computational workflow transparent, reproducible and flexible. Next, I reviewed and benchmarked current methods developed for identifying genes regulated at the level of translation. This allowed me to select a strategy to elucidate that overexpression of two B-cell oncogenes, BCL6 or MYC, is followed by preferential translation of selected transcripts.

Then, I focus on the role of RNA-helicase (DDX3X) in MYC-driven lymphoma. I reveal that that loss-of-function mutations in DDX3X are common in MYC-translocated B-cell lymphomas facilitating early tumour development. By controlling tranlation of selected transcripts, mutated DDX3X buffers the effects of MYC on translation of ribosomal proteins and the rate of global protein synthesis.

Finally, I analyse a large dataset of 79 Ribo-Seq libraries to investigate a genome-wide distribution of translating ribosomes and the extent of non-canonical translation in lymphoid cells. I show pervasive translation of ostensibly non-coding region, and design a knock-down CRISPR screen library to identify those important for B-cell survival.

CHAPTER 2

Materials and methods

2.1 Materials

2.1.1 Overview of genomic sequences and annotations used in this study

Resource	**Source**	**Purpose**
Gencode v.29	Frankish et al. (2019)	Gene and transcript models
GRCh38	Genome Reference Consortium (FASTA file downloaded from Gencode v.29)	Nucleotide sequence of the GRCh38 primary genome assembly
H.sapiens rRNA	RefSeq (FASTA files)	Pre-alignment of Ribo-Seq reads
Reactome	Joshi-Tope et al. (2005)	Pathway knowledgebase
MSigDb	Subramanian et al. (2005)	Molecular Signatures Database for GSEA
sORFdb	Olexiouk et al. (2018)	Proteogenomic database
OpenProt	Brunet et al. (2021)	Proteogenomic database

2.1.2 Overview of external datasets used in this study

Accession ID	Source	Description
GSE125966	McCord et al. (2019a)	RNA-Seq from GOYA clinical trial
GTEx	Consortium et al. (2020)	RNA-Seq samples from 54 tissue sites in non-diseased individuals
TCGA	Weinstein et al. (2013)	Pan-cancer RNA-Seq dataset
CGCI-BLGSP-2019	Grande et al. (2019)	RNA-Seq from Burkitt Lymphoma patients
GSE35163	Schmitz et al. (2012)	RNA-Seq from Burkitt Lymphoma patients
EGAS00001003560	Caeser et al. (2019)	RNA-Seq from Primary Germinal Center B-cells
COSMIC	Tate et al. (2019)	Database of somatic mutations in cancer
MSV000084172	Sarkizova et al. (2020)	Mono-allelic MHC-I peptidome from B721.221 cell line
PXD000332	Deeb et al. (2014)	N-glyco FASP and super-SILAC mass spectrometry from lymphoma patients
PXD002004	Johnston et al. (2018)	TMT 10-plex mass spectrometry from CLL patients and peripheral B-cells
PXD002098	Deeb et al. (2012)	Super-SILAC dataset from lymphoma cell lines
PXD004452	Bekker-Jensen et al. (2017)	Deep-proteome dataset of HeLa cells, tissue samples (colon, prostate, liver) and 5 human cell lines
PXD004746	Khodadoust et al. (2017)	MHC-I and MHC-II peptidomes from lymphoma patients
PXD010808	Khodadoust et al. (2019)	MHC-I peptidome from lymphoma patients

2.1.3 Overview of computational software used in this study

Tool	Source	Purpose
FastQC	Andrews et al. (2017)	Quality check: FASTQ files
Cutadapt	Martin (2011)	Adapter trimming
Bowtie2	Langmead and Salzberg (2012)	Prealignment of Ribo-Seq to rRNA
STAR	Dobin et al. (2013)	Sequencing reads alignment
MultiQC	Ewels et al. (2016)	Summary of QC and alignmnt reports
Samtools	Li et al. (2009b)	BAM files indexing
GenomicFeatures	Lawrence et al. (2013)	Manipulating genomic locations
Samtools	Morgan et al. (2016)	Processing of SAM or BAM files
DESeq2	Love et al. (2014)	Differential expression analysis
edgeR	Robinson et al. (2010)	Differential expression analysis
Deeptools	Ramírez et al. (2016)	Metagene analysis
clusterProfiler	Yu et al. (2012)	GO and GSEA analysis
GATK	Van der Auwera and O'Connor (2020)	SNV calling from RNA-Seq data
Picard	Broad Institute	BAM files processing
GSEA	Subramanian et al. (2005)	GSEA analysis
ORFLine	Hu et al. (2021)	ORF identification from Ribo-Seq data
ORF-RATER	Fields et al. (2015)	ORF identification from Ribo-Seq data
RiboCode	Xiao et al. (2018)	ORF identification from Ribo-Seq data
RibORF	Ji (2018)	ORF identification from Ribo-Seq data
Salomon	Patro et al. (2017)	Isoforms quantification
UCSC toolkit	Kent et al. (2002)	Processing of BigWig files
Chop-chop	Labun et al. (2019)	gRNAs scoring and selection
MaxQuant	Cox and Mann (2008)	Mass spectrometry data analysis
Comet	Eng et al. (2013)	Mass spectrometry data analysis
NewAnce	Chong et al. (2020)	Mass spectrometry data analysis
ProteoWizard	Chambers et al. (2012)	Mass spectrometry data analysis
Xtail	Xiao et al. (2016)	Differential translation analysis
Riborex	Li et al. (2017)	Differential translation analysis
anota	Larsson et al. (2011)	Differential translation analysis
anota2	Oertlin et al. (2019)	Differential translation analysis
deltaTE	Chothani et al. (2019)	Differential translation analysis
Ribowaltz	Lauria et al. (2018)	Ribo-seq QC

2.2 Methods

2.2.1 Next Generation Sequencing library preparation and sequencing

2.2.1.1 RNA-Seq

Performed by Dr. Jie Gao

Total RNA was extracted using NucleoSpin RNA extraction kit (Machery-Nagel, Cat No. 740955.250) according to manufacturer's protocol. 500 ng of total RNA was used to prepare RNA-seq libraries using NEBNext Poly(A) mRNA magnetic isolation module (Cat No. NEB E7490) as per the manufacturer's instruction. Final libraries were amplified by PCR for 12 cycles, purified with AMPure XP beads and analyzed by Agilent Bioanalyser before sequencing on an Illumina Hi-seq4000.

2.2.1.2 Ribo-Seq

Performed by Dr. Jie Gao

Ribosome profiling was conducted as previously described (Ingolia et al., 2012) with minor modifications. 5 million cells per sample were treated with 100 μg/ml of cycloheximide and immediately centrifuged and lysed in 300 μl of buffer containing 20 mM Tris-HCl, pH 7.4, 150 mM NaCl, 5 mM MgCl2, 1% NP40, 1 mM DTT and 100 μg/ml cycloheximide. 100 μl of the lysate were reserved for RNA-Sequencing and the rest treated with DNase I and RNase I. Ribosome monomers were purified using Microspin S-400 columns. The ribosome protected RNA fragments (RPF) were extracted using RNA clean up and concentration kit (23600, Norgen Biotek). RPF were resolved in 15% Novex TBE-Urea gels, stained with SYBR gold and fragments with 26-34 nt were excised from the gel. The RNA was extracted from the gel by electrophoresis using D-tube (MWCO 3.5 kDa, 71506-3, Merck Chemical). Precipitated RNA was dephsphorylated by T4 polynucleotide kinase and ligated to universal miRNA Cloning Linker (NEB) using T4 RNA ligase 2 truncated. cDNA was prepared with SuperScript III and reverse transcription primer containing a degenerate 5-nucleotides molecular barcode sequence. The cDNA was resolved by polyacrylamide gel electrophoresis, excised and extracted by dialysis D-tube (71504-3, Merk). The extracted cDNA was then circularized and PCR amplified. The final library was separated from PCR primers by electrophoresis and extracted by dialysis using D-tube (71504-3, Merk) before sequencing on an Illumina Hi-Seq4000 as 50 nt single-end reads.

2.2.2 Processing and quality control of Next Generation Sequencing data

2.2.2.1 Adapter trimming and alignment to the reference genome

Raw FASTQ files were stripped of adapter sequence using Cutadapt. Reads shorter than 15 nucleotides were discarded. After quality check with FastQC 0.11.5, Ribo-Seq reads were additionally filtered by rRNA using Bowtie2 2.3.4 with seed length 15. The remaining reads were then mapped to the human genome (GRCh38) using STAR 2.5.4a with default parameters. The reference human rRNA index was constructed from RefSeq database. STAR genome index was built with GENCODE v.29 comprehensive gene annotation set.

2.2.2.2 Read counting

Trimmed gene models were built using GENCODE v.29 comprehensive gene annotation set with *GenomicFeatures* R package. The first 30nt and last 30nt of each CDS region were removed to reflect translation elongation intensity, as described previously. Trimmed CDS models representing different transcript isoforms were merged by gene. The number of footprints for 5′UTR and 3′UTR was obtained using a similar logic: the UTRs regions were trimmed by 5 nucleotides adjacent to the annotated start (for 5′UTR) or stop codon (for 3′UTR). As is expected from Ribo-seq, 28-30 nt long fragments with evident triplet nucleotide periodicity relative to start and stop codon were the most abundant and were selected for further analysis. Localisation of ribosomal P-site was determined by offset between 5'end of fragments spanning translation start site and annotated start codon, which was 12 nt for read lengths of 28-30 nt. Per sample per gene P-sites counts matrices were build using *Genomic Ranges* R package allowing for assignment of a read to more than one overlapping features. At least 25 overlapped bases were required to assign a read to a gene. Corresponding RNA-Seq samples were counted using the same gene models. Genes with low counts were filtered out with a threshold of minimum 128 counts

2.2.3 Differential translation analysis of Ribo-Seq data

Differential translation analysis was performed as previously described (Sendoel et al., 2017; Hsieh et al., 2012). Briefly, differentially expressed genes were identified using DESeq2. RNA-Seq and Ribo-Seq derived read counts were analysed separately. Log_2 fold change in translation efficiency (TE) was computed by subtracting Log_2 fold change in mRNA from log_2 fold change in ribosome footprints abundance. Both values were obtained from the DESeq2 output.

The following rules were applied to define differentially translated genes:

1. Statistically significant change in ribosomal footprints abundance (evaluated with standard DESeq2 workflow), FDR < 0.1,

2. Absolute mean $\log_2$ fold change in mRNA abundance < 0.3,

3. Absolute mean $\log_2$ fold change in TE abundance > 0.3.

In addition, two other gene regulatory classes were identified dependent on the relationship between Ribo-Seq and RNA-Seq measured expression changes. For genes with statistically significant change in mRNA abundance (FDR < 0.1 and absolute $\log_2$ fold change in mRNA abundance > 0.3) and concordant change in Ribo-Seq (mean $\log_2$ fold change in ribosomal footprints abundance > 0 or < 0 depending on the direction of change and regardless of the statistical significance) are classified as 'Homodirectional'. When the direction of change is opposite, the genes are classified as either 'buffered mRNA up' or 'buffered mRNA down'.

2.2.4 Differential expression and downstream analysis of RNA-Seq data

2.2.5 Metagene analysis of iCLIP and Ribo-Seq data

A metagene analysis for scaled density of Ribo-Seq reads or iCLIP hits relative to start and stop codon was performed using deeptools. The coverage of sequencing reads was normalized per sample by the total number of uniquely mapped reads (CPM) excluding sex chromosomes. Scaled coverage per each transcript was computed using computeMatrix function with parameters as follows: `scale-regions -m 10000 -bs 20 -m 5000 -b 3000 -a 3000 -p 40 --metagene --exonID CDS --transcriptID transcript --skipZeros`. Transcripts models were built using GENCODE v.29 basic gene annotation set.

2.2.6 Differential expression analysis of RNA-Seq data

Uniquely mapped reads were assigned to genes using featureCounts function from Rsubread package allowing for assignment of a read to more than one overlapping features. At least 25 overlapped bases were required to assign a read to a gene. Differential expression analysis was performed using standard workflow from DESeq2 package.

2.2.7 Downstream data analysis

Identified up-regulated and down-regulated genes were used to perform gene ontology analysis with enrichGO function from clusterProfiler package. Enrichment scores were

computed using in-house scripts by taking the ratio between the number of differentially expressed genes overlapping with a gene ontology set and the number of background genes assigned to this gene set. Pathway analysis was performed using browser-based Reactome Pathway Database. A list of all expressed genes detected in RNA-Seq was used as a background set for over-representation testing. Gene Set Enrichment Analysis (GSEA) was performed using the GSEA function from clusterProfiler package. Gene expression measurements were normalized using variance stabilising transformation, as implemented in DESeq2 package, and analysed for enrichment in hallmark gene set from MSigDB v. 7.0

2.2.7.1 Individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP)

iCLIP libraries were performed by Dr Chun Gong in the Hodson lab using two lymphoma cell lines as well as primary human germinal centre B cells purified from donor tonsil.

iCLIP data preprocessing (alignment, trimming and peaks calling) was performed by Dr Igor Ruiz De Los Mozos using iMaps web server. The downstream analysis and data visualisation was performed by me. Briefly, Unique Molecular Identifiers were used to distinguish and remove PCR duplicates before removing experimental barcodes and Solexa adapters. The trimmed reads were mapped to GENCODE GRCh38 v.29 using STAR with default parameters. First nucleotide after the UMI was assigned as the crosslink site defined by the truncated cDNA. Crosslink significant sites were determined by the iCount peaks finding algorithm (False Discovery Rate < 0.05), by weighting the enrichment of crosslinks versus shuffled random positions. Neighbouring cDNA start position less than 15 nt apart were joined to form high confidence crosslink clusters with iCount clusters function. Genes with more than 4 cross-linking peaks in at least one experiment were considered as valid DDX3X targets.

2.2.8 Identification of DDX3X mutations from RNA-Seq data

I used published RNA-Seq data to identify cases with mutation of the DDX3X gene. Single Nucleotide Variant (SNV) calling was performed according to GATK guideline using paired-end RNA-Seq data from diagnostic biopsies of 553 patients enrolled in the GOYA trial - a phase III of 1st line treatment in DLBCL. Briefly, RNA-Seq paired-end reads were mapped to the reference genome using STAR 2.5.4a 2-pass mode (Dobin et al., 2013). Read groups were added with AddOrReplaceReadGroups and duplicated reads were identified with MarkDuplicates script from Picard tools. Sequences overhanging intronic regions were hardclipped and STAR mapping qualities were reassigned to match GATK software. Variant calling was performed with HaplotypeCaller GATK script, with phredScore 20 as minimal variant calling confidence. Variants clusters (at least 3 valid

variants in a window of 35 bases) were removed to diminish the effect of RNA-Seq mapping errors. Standard variants quality filtering was applied with VariantFiltration GATK script: Fisher Strand values > 30.0 and Qual By Depth < 2.0 with. Individual SNVs were then annotated with gene names and their predicted consequences on protein function using VariantAnnotation Bioconductor package and gene models from GENCODE comprehensive gene annotations set (v.28). In order to identify samples with potential DDX3X mutation, ENSG00000215301 (gene ID) and ENST00000644876 (transcript ID) DDX3X models were used as a reference. All nonsense, frameshift and non-synonymous SNVs with the ratio of variant coverage : reference coverage > 0.2, localised in DDX3X helicase domain and not previously reported in the ExAC database of common population variants were considered as valid hits.

2.2.9 DLBCL Cell-of-origin identification from RNA-Seq data

Classification of DLBCL biopsies into four transcriptomic subtypes: ABC, GCB, Unclassified and Molecular High Grade (MHG) DLBCL was performed as previously described (Reddy et al., 2017; Sha et al., 2019).

2.2.10 Chromosome Y expression identification from RNA-Seq data

Chromosome Y expression identification in human RNA-Seq data was performed using decision tree algorithm implemented in rpart R package. The model was trained using RNA-Seq data from Genotype-Tissue Expression (GTEx) Project, which includes gene expression samples from 54 tissue sites in non-diseased individuals (11688 samples in total)(2013). The GTEx datasets (V8) were obtained from https://gtexportal.org/home/. Raw RNA-Seq counts were filtered and TMM normalised. Per gene scaled gene expression values were used as an input. The GTEx data were randomly split into a training and test set comprising, respectively, 80% (9333 samples) and 20% (2355) of the data. The algorithm running on default parameters achieved high performance: F1 score = 0.9973, AUC = 0.9973, 8 males were misclassified. *KDM5D*, *DDX3Y*, *USP9Y*, *RPS4Y1*, *TXLNGY*, *XIST* were identified as classifying genes. In order to assess the ability of the algorithm to classify cancer samples, GTEx trained model was benchmarked against cancer dataset. TCGA gene expression data (RNA-Seq only)(Weinstein et al., 2013) were downloaded using TCGAbiolinks, the Bioconductor package for integrative analysis with GDC data. Similarly, the dataset was split into a training (80%, 9198 samples) and test set (20%, 2336 samples). The algorithm achieved lower performance than in GTEx data: F1 score = 0.9624, AUC = 0.9779, 46 males and 37 females were misclassified on default parameters. When the GTEx trained model was tested on the TCGA test dataset, the number of

misclassified females and males was 2 and 295, respectively. The 295 males classified as females showed remarkably lower expression of genes localised on chromosome Y, which could reflect previously reported loss of chromosome Y during oncogenesis (Dunford et al., 2017). In order to obtain high-confidence set of male DLBCL patient samples with chromosome Y expression, the GTEx trained model was used to classify samples in the GOYA dataset.

2.2.11 Hierarchical *de novo* identification of translated regions from Ribo-Seq data

Ribo-Seq and RNA-Seq FASTQ files were processed and a standard quality check was performed as described earlier in section 2.2.2. P-site location was determined automatically with Plastid suite using `psite` and `phase_by_size` functions with default settings. Only reads of length between 25 and 33, with clear three nucleotide pattern, were used to call ORFs. Aligned sequencing reads were used to run ORFLine (Hu et al., 2021), ORF-RATER (Fields et al., 2015), RibORF (Ji, 2018) and RiboCode (Xiao et al., 2018) with default settings. Harringtonine treated samples, if available, were used as an additional input for ORF-RATER run to increase precision of locating translation initation sites. Identified ORFs were annotated over the reference transcriptome that was customised for lymphoid cells, so that it contains only one the most highly expressed representative isoform per gene. Three different RNA-Seq datasets were queried for that purpose: 1) 50 base pair single-end RNA-seq data matching Ribo-Seq dataset (77 samples), 2) 50 base pair paired end RNA-Seq data from GOYA clinical trial (553 samples) (McCord et al., 2019a). Transcript-level quantification of expression was performed with Salmon following the recommendations from the user manual (Patro et al., 2017). Salmon *.sf* files were loaded to R with listings package. Scaled TPMs (transcript expression abundance scaled to library size) were used to determine a mean expression of each isoform. The isoform with the highest mean expression was selected as representative. For protein coding transcripts, only protein coding isoforms were considered.

Identified ORFs were classified as canonical if both, start and stop codon were present in Gencode V.29 annotations. All the remaining ORFs were assigned to one out of seven types as defined in **Table 2.1**.

Table 2.1: Unified naming system for classifying identified ORFs

Unified type	Definition
canonical	Matching start and stop codon
uORF	Initiating and terminating downstream of the canonical ORF.
dORF	Initiating and terminating upstream of the canonical ORF.
novel	Localised in non-coding transcript.
internal	Contained within known CDS but out-of-frame with canonical product.
truncation	Matching stop codon but initiating downstream of the canonical start site.
extension	Matching stop codon but initiating upstream of the canonical start site.
readthrough	Matching start codon but terminating downstream of the canonical termination site.
overlap uORF	Initiating upstream of the canonical start codon and terminating upstream of the canonical termination site.
overlap dORF	Initiating upstream of the canonical termination site but terminating downstream of the canonical termination site.

2.2.12 Reanalysis of published mass spectrometry datasets

ProteomeXchange database (Deutsch et al., 2016) was searched for proteomic experiments performed in lymphoid cells (keywords: B-cells, lymphoma, lymphocyte) covering variety of experimental conditions and mass spectrometry (MS) techniques. Experiments with incomplete submission, targeted experiments (eg. co-immunoprecipitation of a single protein) or with very few replicates were excluded. RAW files were downloaded from PRIDE (Perez-Riverol et al., 2019) or MassIVE repository.

To avoid forced misanotation of MS/MS spectra to noncanonical products, a customised reference FASTA file was build by merging a database of human canonical proteins downloaded from UniProt with amino-acid sequences of predicted micropeptides. To decrease search space noncanonical ORFs shorter than 6 amino-acids and showing more than 20 % of in-frame overlap with known ORFs were filtered out.

Andromeda search engine implemented in MaxQuant 1.6.3.4 software (Cox and Mann, 2008) was used to process all collected MS datasets. The parameters were set as follows: precursor mass tolerance 20 ppm, MS/MS fragment tolerance 0.5 Da; PSM FDR and Protein FDR threshold 1%. Noncanonical reference sequences were provided as *proteogenomic fasta file*, that allows for separate FDR calculations for canonical and noncanonical peptides. For immunopepditomics data, NewAnce workflow was followed (Chong et al., 2020), that required additional search with Comet engine (Eng et al., 2013). To match Comet input requirements, RAW files were converted to mzXML format with the MsConvert from ProteoWizard tool (Chambers et al., 2012). Search parameters for immunopeptidomics data were set as described above with small modification: peptide length of 8–15 amino-acids was set when searching MHC I data and 8-25 for MHC II. Variable and fixed modifications were kept as in the original publication.

2.2.13 Analysis of proteogenomic data downloaded from OpenProt and sORFdb

FASTA files containing the sequence of peptides deposited in sORFdb and OpenProt 1.6 databases were downloaded from `https://www.openprot.org` (Brunet et al., 2021) and `http://sorfs.org` (Olexiouk et al., 2018), respectively. Amino-acid sequences of candidate ORFs identified in this study were queried against those with the databases using BLASTP (Altschul et al., 1990) with default settings. ORFs matching the queried database in at least 95% were defined as valid matches.

2.2.14 Evolutionary conservation of identified ORFs

Evolutionary conservation of each identified ORF was evaluated with PhastCons score, which represents the probability that each nucleotide belongs to an evolutionary conserved element. Per base values for GRCh38/hg38 were downloaded from UCSC. BigWig files obtained from three multiple alignments were selected: 1) 100 way: representing conservation between 99 vertebrate genomes and the human genome 2) 20 way: 29 vertebrates genomes, and 3) 7 way: 6 vertebrate genomes, including 3 primates.

Averaged scores over ORF regions were computed with a command line tool *bigWigAverageOverBed* from UCSC toolkit (Kent et al., 2010).

2.2.15 CRISPR screen design

To identify ORFs fit for CRISPR screening, entries fulfilling the following criteria were removed: 1) ORFs shorter than 30 nucleotides, 2) with mean expression below 20th percentile, 3) sharing more than 50% of sequence with known protein coding region, and 4) having a unique part (not overlapping any CDS) shorter than 30 nucleotides. The remaining ORFs were collapsed by their genomic coordinated, extended 16 nucleotides in $5'$ direction and used to generate a BED file. Because CRISPR-Cas9 enzyme cuts 3-4 nucleotides upstream of the PAM sequence, extension of the queried region allows to maximise the number of candidate gRNAs targeting the translation start site. BED file was used as an input for Chop Chop, a Python tool to search for all gRNA given genomic coordinates (Labun et al., 2019). A customised wrapper script was used to search for all regions at once. Chop chop parameters were set as follows:

```
chopchop.py -T 1 -M NGG --maxMismatches 3 -g 20
--scoringMethod ALL -f NN -n N -G hg38 -o temp
-t WHOLE -Target [REGION]
```

Obtained gRNAs were then filtered based on their GC content and the number of predicted off-target locations. All gRNAs with any predicted off-targets with 0 mismatches

or having GC content lower than 40% or higher than 70% were removed. gRNAs were reassigned back to ORFs and ORFs with fewer than 5 gRNAs were considered untargetable and filtered out. All sequences were prepended with a G at 5′ end to facilitate transcription from the U6 promoter. gRNAs were scored as implemented in the CHOP-CHOP workflow. gRNAs with a G at position 20 (upstream of PAM were prioritised) and on-target activity scores were computed as described in (Moreno-Mateos et al., 2015; Xu et al., 2015; Doench et al., 2014; Dutt et al., 2011).

Remaining ORFs were then scored and ranked based on the features increasing their biological importance. Following characteristics were used to compute the mean score per ORF:

1. Mean scaled ORF score (Bazzini et al., 2014): reflecting the accumulation of ribosomal footprints in the first reading frame: $ORF_{score} = log_2(RRF_1 + RRF_2 + RRF_3)$, where RRF is the proportion of reads in given reading frame; scaled so that it takes values from 0 to 1, when 1 indicates a situation when all ribosomal footprints belong to the first reading frame and 0 means no first frame preference.
2. Expression level: measured as percentile of mean expression across all unmerged samples included in the study.
3. Identification in the reanalysed mass spectrometry studies: 2 points for ORFs identified in both immunopeptidomics and full proteome studies, 1 - identified in one and 0 - no found in proteomics.
4. Present in Uniprot or any of the two proteogenomic databases (sORFdb or OpenProt): 1 - if present, 0 - if not.
5. Mean conservation score computed by averaging three phasCons scores (100way, 20way and 7way) computed as described above in section 2.2.14, takes values from 0 (no evolutionary conservation) to 1 (full evolutionary conservation).
6. The number of ORF finding tools that has identified an ORF as actively translated scaled, so that it takes values from 0.25 (detected by 1 tool) to 1 (detected by all tools)
7. The number of samples in which an ORF was detected by at least one tool scaled so it takes values from 0.0084 (detected in 1 sample) to 1 (detected in all samples)
8. The proportion of differential expression analyses, where an ORF reached statistical significance threshold.

The library was constructed to target top scoring ORFs with 4-5 top scoring gRNAs per each and synthesised as a pool of 6000 single-stranded oligos (Twist Bioscience).

2.2.16 Figures preparation

Figures were prepared using R package ggplot2 (Wickham, 2016) and composed into panels with Adobe Illustrator or Adobe Photoshop.

2.2.17 R and Bioconductor

Statistical analysis was performed with R 4.0.2. The version of Bioconductor was 3.12.

CHAPTER 3

Genome wide quantification of translation in lymphoid malignancies

3.1 Background

Deregulation of protein synthesis is a hallmark of malignant transformation, however the molecular mechanisms underlying those processes still remain elusive. The transcription factors BCL6 and MYC are essential for induction and maintenance of GC reaction (De Silva and Klein, 2015) and, when deregulated, become drivers of lymphomagenesis (Basso and Dalla-Favera, 2015). During the GC reaction, both proteins drive extensive reprogramming of B-cell gene expression landscape, which has been studied mainly at the level of transcription. The contribution of other levels, including post-transcriptional regulation, is poorly understood. Sendoel et al. (2017) showed that overexpression of SOX2 in skin epidermis is followed by broad changes in translation programme. SOX2 is a transcription factor that is highly expressed in fully developed squamous cell carcinoma as well as in cells considered to be *cancer stem cells* for this type of tumour. I hypothesised that similar reprogramming of protein synthesis may follow overexpression of the transcription factors BCL6 and MYC in GC B-cells.

We took advantage of a primary GC B-cells culture system, that has been developed in the Hodson lab (Caeser et al., 2019). Primary GC B-cells were isolated from tonsil tissue discarded after elective tonsillectomy and cultured with Follicular Dendritic Cells (FDCs) expressing CD40L and IL 21, which mimic the supportive role of GC microenvironment. The co-culture system has been described in detail by Caeser et al. (2019). To test whether BCL6 and MYC, in addition to their transcriptional effects, are responsible for changes at the level of translation, we decided to perform Ribo-Seq and RNA-Seq experiments in

primary GC B-cells overexpressing either MYC-t2A-BCL2 or BCL6-t2A-BCL2 or BCL2 alone **(Figure 3.1 A)**. The co-expression of BCL2 and BCL6 or MYC from the same viral construct with T2A linker was essential to keep B-cells alive in culture and perform the experiments **(Figure 3.1 B)**. Transduction of either MYC-t2A-BCL2 or BCL6-t2A-BCL2 is associated with malignant transformation of the primary GC B-cells inducing their prolonged expansion and survival (Caeser et al., 2019).

Cell culture, transduction and sequencing libraries were prepared by Dr. Jie Gao as described in section 2.2.1.1.

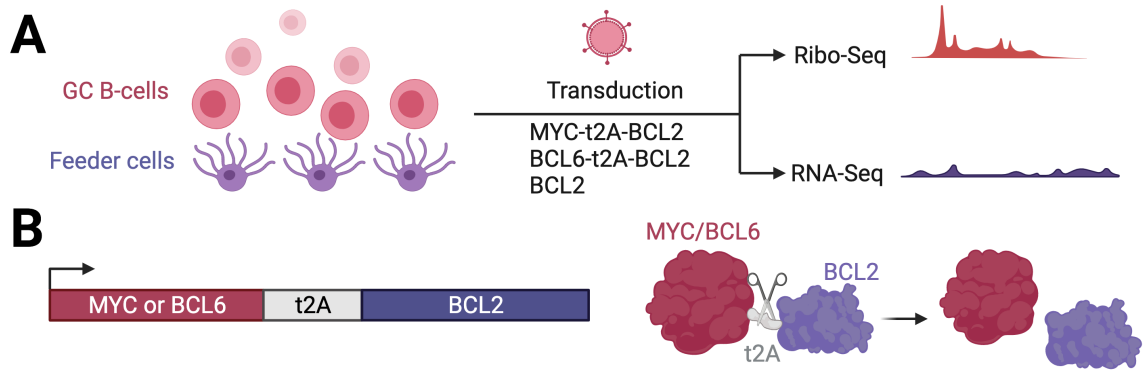

Figure 3.1: Studying translational regulation in primary GC B-cells overexpressing common lymphoma oncogenes

A) Flowchart showing the experimental design for investigating translational changes following BCL6 or MYC overexpression. Primary GC B-cells are cultured with a monolayer of immortalised Folicular Dendritic Cells-like feeder cells. The feeders express CD40L and IL 21 providing the essential survival signal for cultured primary cells.

B) Diagram illustrating the design of a viral construct for simultaneous expression of BCL2 with BCL6 or MYC using t2A linker. Thosea asigna virus 2A (t2A) is a small peptide (19-22 amino acids), which undergoes self-cleavage resulting in simultaneous expression of two or more proteins in equivalent amounts.

In this chapter:

1. I introduce RiboStream, a bioinformatic pipeline developed to process Ribo-Seq and RNA-Seq data simultaneously,

2. I provide an overview and comparison of available tools for identifying differentially translated genes. Despite growing popularity of Ribo-Seq in studies on protein translation, the analytic workflow is not well established. I use an external dataset of 54 Ribo-Seq and RNA-Seq samples to evaluate the available tools and choose the approach for further analysis,

3. I apply my optimised strategy to analyse changes in the translation as non-malignant GC B-cells undergo transformation driven by 2 different oncogene combinations.

3.1.1 Establishing a bioinformatic pipeline for processing of translatome and transcriptome data

The analysis of large genomic datasets usually involves sequential application of various command line tools, software packages and custom scripts. The design of such workflows depends on the exact next-generation sequencing technology and is usually tailored to available computational resources. In order to facilitate processing of dataset for this and future projects, there was a need to develop an *in-house* computational pipeline.

I developed a pipeline, which I named RiboStream, for parallel processing of Ribo-Seq and RNA-Seq datasets; `https://github.com/ashakru/RiboStream_bpipe.git`. Subsequent stages of the workflow were linked using Bpipe, a Groovy language platform for managing bioinformatic jobs (Sadedin et al., 2012). A workflow management system, such as Bpipe, can aid to achieve reproducibility, and transparency of the analysis. Bpipe pipeline architecture has numerous advantages over running all the jobs manually or through a series of Shell scripts.

The central idea of a workflow management system is to automate, orchestrate and monitor the execution of subsequent stages of the data processing. Automatic log of executed commands and the outputs management increase readability and control on the workflow, that facilitate troubleshooting and customisation, if needed.

It is not surprising that workflow management systems have been widely adopted by the bioinformatic community and they are now considered a part of a good research practice. The choice of Bpipe has been dictated by the simplicity of Bpipe architecture, compatibility with a cluster resource manager, and the avaliability of comprehensive documentation and users community support.

Briefly, RiboStream takes raw FASTQ files with Ribo-Seq or RNA-Seq reads, as received from a sequencing facility, and returns: aligned BAM files, various quality check metrics and read count matrices, which ready for downstream analysis tailored to the biological question of interest. Optionally, Sequence Read Archive (SRA) accession numbers can be provided instead of FASTQ files. In such scenario the pipeline downloads FASTQ files from SRA, the largest publicly available repository of next generation sequencing data. Flexibility and control over each stage is provided by a single configuration file, where all the parameters of executed tools can be adjusted. Sample-specific parameters, such as the localisation of the FASTQ files (or SRA accession numbers), type of the experiment (Ribo-Seq or RNA-Seq) or the sequence of the adapters for trimming, are contained in a sample sheet and parsed by the pipeline when needed. An outline of the pipeline is shown in **Figure 3.2**

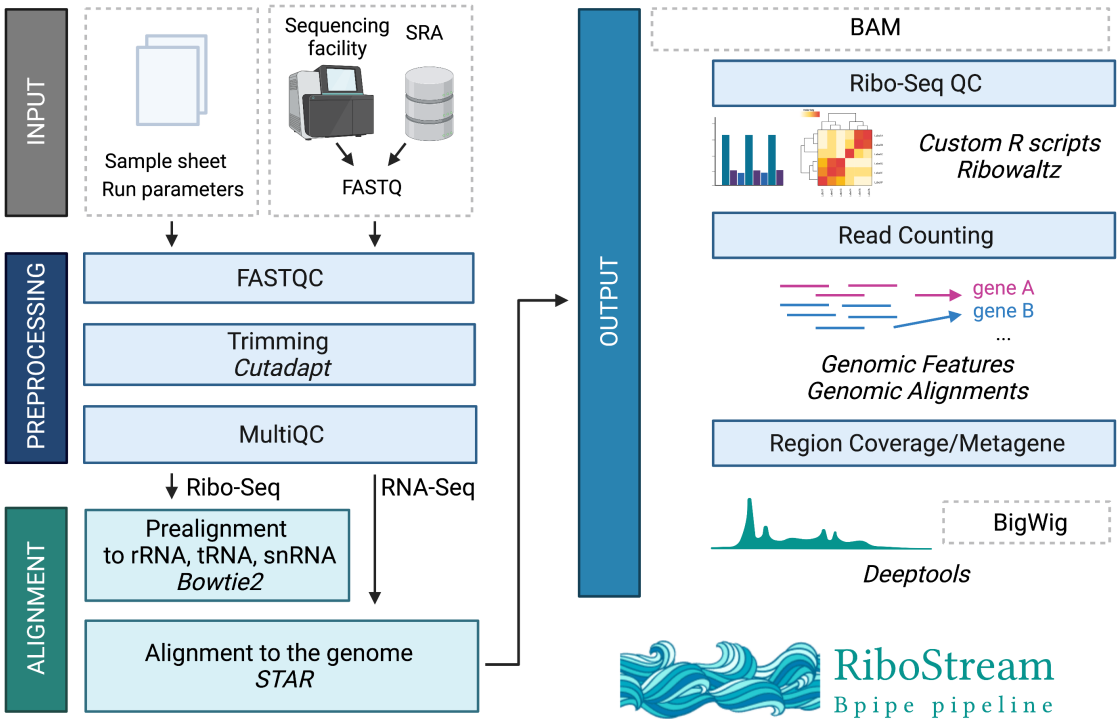

Figure 3.2: Overview of RiboStream, a Bpipe pipeline for Ribo-Seq data analysis
Flowchart showing a basic bioinformatic workflow that I developed for Ribo-Seq and RNA-Seq data processing. A typical run performs automatic preprocessing of raw FASTQ files, alignment to the reference genome and routine quality check focused on ribosomal footprints data. The output, in a form of count matrices, BAM and BigWig files, is ready for a downstream analysis. Bioinformatic tools implemented in the workflow at each stage are indicated.
SRA - Sequence Read Archive; QC - quality check

While most of the stages are the same for both data types, there are three steps specific and for Ribo-Seq samples: prealignment to non-coding RNA sequences, extended quality control and modified reads counting strategy.

The first involves removal of common contaminating sequences derived from ribosomal RNA (rRNA), tRNA or small nuclear RNA (Ingolia et al., 2012). The advantage of this extra data purification step is an increase in the proportion of true ribosome-derived mRNA footprints in the final BAM file, containing reads aligned to the human genome. This facilitates the interpretation of the alignment pattern and improves the estimation of the ribosomal P-site location. An additional benefit is a smaller size of the final output file, that accelerates further processing.

Quality control is a critical step of all next generation sequencing experiments and usually involves evaluation of raw reads parameters (e.g. total number of reads, base quality scores, GC content), alignment efficiency and reproducibility of biological or technical replicates. In addition there are several metrics characterising a good quality Ribo-Seq

experiment, that should be checked before performing further analyses. Ribo-Seq specific quality measures include the proportion of uniquely mapped reads mapping to various genomic locations (introns, exons, known CDS or untranslated regions), and transcript biotypes (e.g. protein coding, lncRNAs, pseudogenes). We anticipate the majority of ribosomal footprints to map to CDS regions of known protein-coding genes. This step was performed with *Ribowaltz* package and custom R scripts utilising basic functions of the versatile *GenomicFeatures* and *GenomicAlignments* packages. Another important quality check is the pattern of 5′ends of the footprints around annotated start and stop codons. This should reveal enrichment of ribosomal footprints in CDS comparing to 5′UTR or 3′UTR and clear three-nucleotide periodicity of alignment reflecting single frame preference of translated ribosomes **(Figure 3.3 A)**. These patterns usually are evident for reads of length between 26 and 30 nucleotides, which most likely correspond to true ribosomal footprints.

A similar metric, which also investigates the characteristic distribution of the footprints, is a metagene profile. RiboStream performs a metagene analysis to examine the distribution of aligned Ribo-Seq reads over all known protein-coding transcripts. The aggregated (metagene) profile visualises general patterns in read coverage over sets of genomic regions of interest, for example promoters, transcripts or CDS **(Figure 3.3 B)**. A metagene analysis examining the density of Ribo-Seq reads mapping between known start and stop codons is vital to ensure data quality and may reveal biologically important changes in ribosome occupation between experimental conditions (Gerashchenko and Gladyshev, 2014). The expected pattern characterising a successful Ribo-Seq experiment is the footprint density at the first 30 to 40 codons that is far greater than in the rest of CDS. In the literature the mechanism of this pattern was widely discussed. One proposed biological explanation is that it reflects the speed of translation initation, which is slower than translation elongation. Specific codon context around the start codon may explain the accumulation of the ribosomal footprints mirroring local translation rate (Ingolia et al., 2009; Tuller et al., 2010).

An alternative hypothesis involves treatment with translation inhibitors. Cycloheximide, which was used in this and numerous other studies including the original Ribo-Seq paper (Ingolia et al., 2009), inhibits translation elongation by interrupting ribosome translocation. By the time all cells are saturated with the inhibitor, some ribosomes keep initiating, which may explain the observed accumulation of footprints in the first codons (Gerashchenko and Gladyshev, 2014).

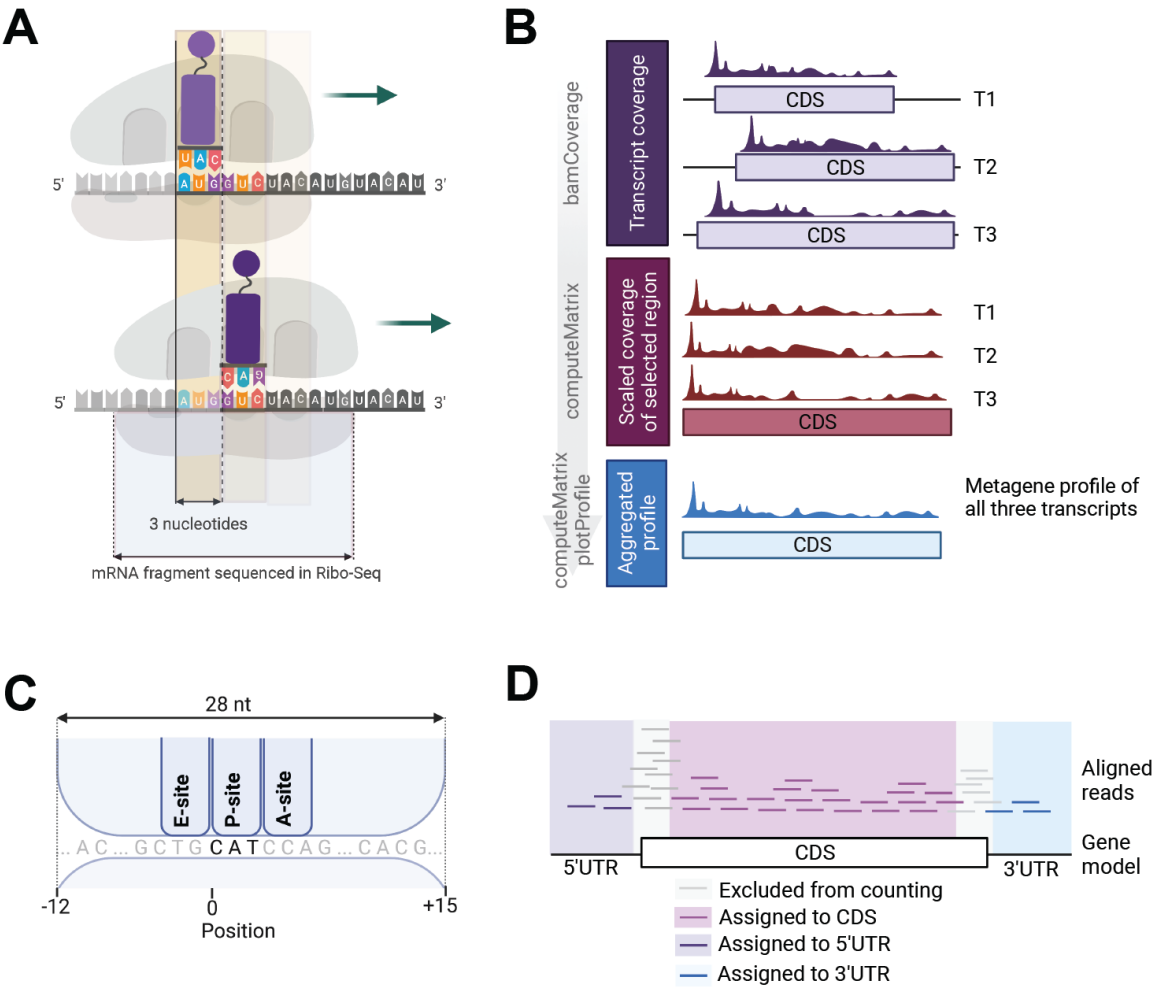

Figure 3.3: Illustration of critical steps in data processing specific to Ribo-Seq datasets

A) Diagram showing the mechanism explaining periodical pattern of ribosome footprints alignment, which suggest involvement in active translation.

B) Flowchart of metagene analysis implemented in RiboStream. The workflow consists of three steps in which read coverage profiles are scaled to the same length, and then aggregated to create a metagene profile. The final plot represents an average mapping density over all known CDS. *Deeptools* commands (bamCoverage, computeMatrix or plotProfile) used at each stage are indicated.

C) Diagram showing the position of estimated ribosomal P-site in a 28 nucleotide long Ribo-Seq read.

D) Diagram showing the strategy to assign mapped reads to different genomic regions: 5′UTR, CDS or 3′UTR.

CDS - Coding Sequence Region, 5′UTR - 5′ Untranslated Region, 3′UTR - 3′ Untranstaled Region

An initiation peak also occurs in untreated Ribo-Seq samples (Ingolia et al., 2009; Weinberg et al., 2016), but they are much smaller than the peaks in samples treated with translation inhibitor. Small peaks can be also observed at the stop codon, but the mechanism of this is unknown. Regardless, the mechanism, the accumulation of Ribo-Seq reads around the start codon has real consequences for the data interpretation. Firstly, it is a useful quality measure of the purity of Ribo-Seq signal. Secondly, it allows to identify translation initiation sites directly from the data, which is utilised by various ORF identification algorithms (Calviello and Ohler, 2017). Lastly, this requires a different read counting strategy to the approach established for RNA-Seq data.

Similairly, as differential expression analysis is a core application of RNA-Seq technique, differential translation is one of the main aims of a typical Ribo-Seq experiment. Counting how many fragments have aligned to each gene is an essential step of such analysis. Ribo-Seq read counting in the Bpipe pipeline was implemented to follow the original protocol from Ingolia et al. (2009) with minor modifications. Ribosomal footprints were assigned to genomic regions (CDS, intron, 5′UTR or 3′UTR) based on the position of the estimated P-site **(Figure 3.3 C)**. I built the models of the genomic regions with *GenomicFeatures* R package using GENCODE v.29 comprehensive gene annotation set as reference. The first 15 and last 15 nucleotides of each CDS region were trimmed, so that the assigned footprints reflect the baseline translation intensity rather than the rate of initiation or termination (Figure 3.3 D). Trimmed CDS models representing different transcript isoforms were merged by gene. Only fragments with evident triplet nucleotide periodicity relative to start and stop codon, typically between 27 and 30 nucleotides long, were selected. The pipeline determines the position of the ribosomal P-site with *psite* script from Plastid toolkit that looks at the offset between 5′ end of fragments spanning translation start site and annotated start codon, which, usually, is 12 nucleotides for reads in the range between 27 and 30 nucleotides **(Figure 3.3 C)**. Finally, per sample and per gene counts matrices are built using *Genomic Ranges* R package allowing for assignment of a read to more than one overlapping features **(Figure 3.3 D)**. Corresponding RNA-Seq samples are counted using the same gene models.

Processing of a small Ribo-Seq experiment (9-12 samples) with RiboStream installed on High Performance Computing Cluster (18 nodes, 176 CPUs) takes about 3 hours.

3.1.2 Quality of translatome profiling in primary GC B-cells

Overall the average number of uniquely mapped reads in Ribo-Seq samples was almost 80 million reads of which about 70% corresponded to rRNA sequences **(Figure 3.4 A)**. Although the amount of rRNA-mapping reads depends on the organisms, footprint isolation method, and translational status, they usually account for the majority of sequenced reads (McGlincy and Ingolia, 2017). Given the fact that a ribosome-footprint complex consists of several kilobases of rRNA but only about 28 bases of ribosome protected mRNA, this is not unexpected (Ingolia et al., 2012). Higher rRNA content may also be observed in conditions associated with global change in translation rate, when the ratio of actively translated ribosomes to other RNAs is affected (Ingolia et al., 2012). In contrast to a routine rRNA-depletion library strategy for RNA-Seq experiment, only a few specific rRNA fragments account for the large proportion of contamination observed in Ribo-Seq. This may be related to the RNAse digestion of rRNA at specific and reproducible positions (Ingolia et al., 2012). Our study adopted the rRNA removal protocol introduced in the original Ribo-Seq workflow, where contaminating sequences are hybridised to biotinylated oligonucleotides and depleted with streptavidin beads (Ingolia et al., 2012).

Of non-rRNA reads, the majority was about 29 nucleotides long corresponding to the length of the ribosome protected mRNA fragments. The average GC content of uniquely mapped reads was about 50 %. In about 60% mapping destination was known coding sequence region. **(Figure 3.4 B-C)**.

The analysis of a mean distribution of the reads 5′ end around the start and stop codons, revealed that the 5′ end starts aligning about 12 nucleotides upstream from the annotated start codon and shows strong 3-nucleotide periodicity **(Figure 3.5 A)**. The strongest 3 nucleotide pattern of alignment was seen for reads between 26 and 29 nucleotides long, which is in line with previous work. (Ingolia et al., 2009, 2012). The 12 nucleotide offset between 5′ end of a read and TIS allowed to estimate the position of the ribosome P-site, as was described in the section 3.1.1. The distribution of estimated P-sites around start codon presented clear frame preference and expected accumulation in the first 20 bases of the CDS **(Figure 3.5 B)**. The same analysis applied to samples treated with different translation inhibitor, harringtonine, which immobilises ribosomes shortly after translation initiation, revealed distinct pattern. While samples treated with cycloheximide showed evident 3′nucleotide periodicity over the entire length of CDS, estimated P-sites from the harringtonine group were found almost exclusively in the first 25 nucleotides **(Figure 3.5 B)**.

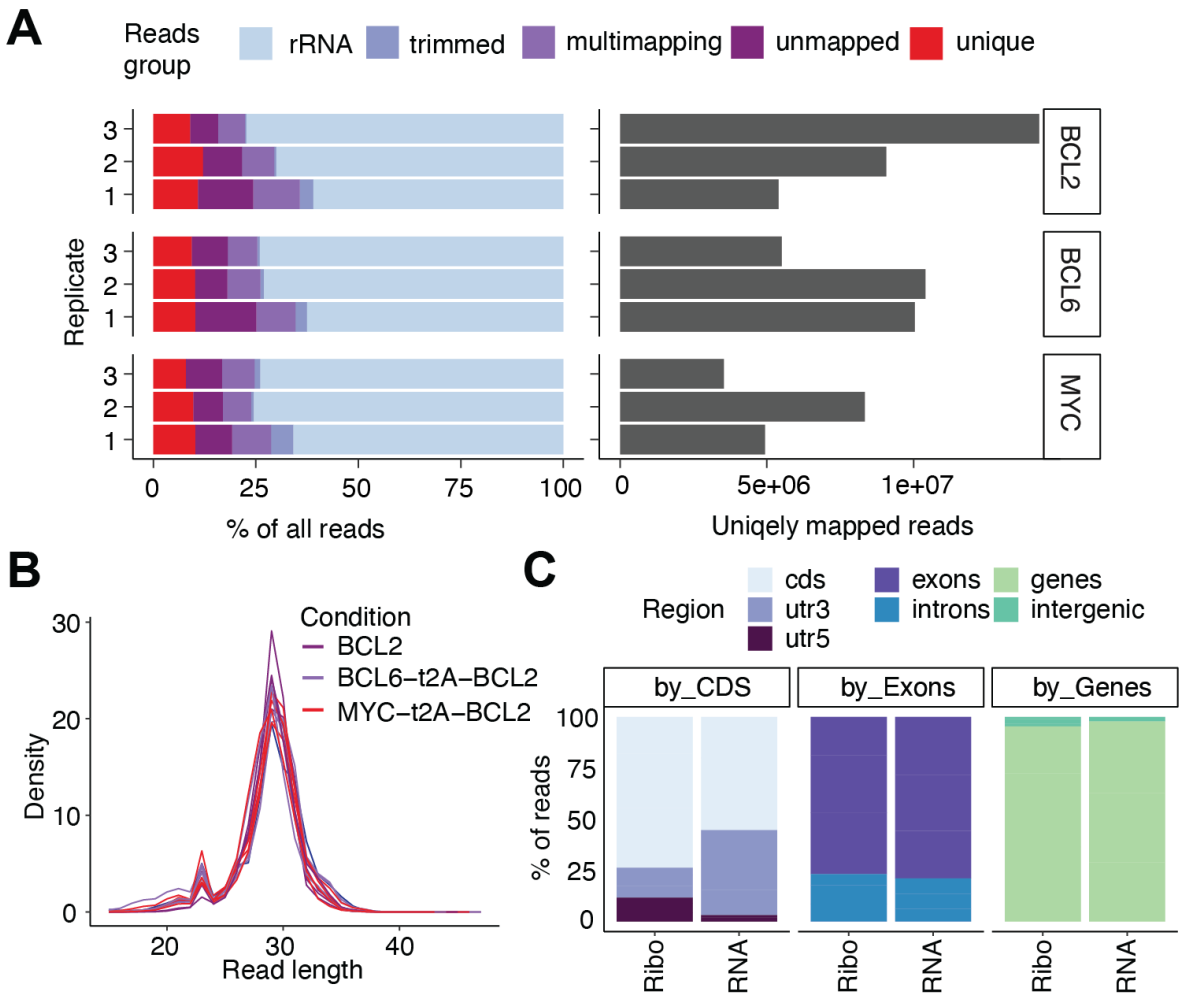

Figure 3.4: Basic mapping statistics metrics for Ribo-Seq reads for the BCL6/MYC experiment in human GC B-cells

A) The fate of Ribo-Seq reads after the two rounds of alignment shown for each replicate. The left panel shows the percentage of reads in each of the five group (aligned to rRNA, removed during trimming stage, mapped to multiple locations, unmapped or aligned uniquely). The right panel depicts the total number of uniquely mapped reads that was eligible for the further analysis.

B) Histogram of read length of uniquely mapped Ribo-Seq reads stratified by experimental condition.

C) Distribution of mapped reads from Ribo-Seq and RNA-Seq experiments to gene features, showing the expected restriction of Ribo-Seq reads to the CDS with only a small portion mapping to UTRs.

On a global scale, a metagene plot of cycloheximide treated samples showed a characteristic peak at the expected translation initation site (TIS) and abrupt drop-off the signal after translation termination site (TTS). The pattern of ribosomes occupancy was similar for all experimental conditions **(Figure 3.5 C)**. When P-sites were stratified by transcript regions (5′UTR, CDS or 3′UTR) and read length, they, again, showed a strong frame restriction in known CDS regions **(Figure 3.5 D)**.

Even in a good quality experiment, genes with low expression level have typically higher variance which may decrease the sensitivity of differential expression analysis. With only few read counts assigned, it is difficult to distinguish biological differences from technical (sampling) noise. Therefore, a common practice is to remove those lowly expressed genes before attempting differential expression analysis. I estimated the threshold of this filtering adopting the strategy introduced by Ingolia et al. (2009), removing all genes with number of RNA-Seq reads lower than 128 in more than 25% of samples. I also filtered out histone genes, which are not polyadenylated and therefore underrepresented in our RNA-Seq data with poly(A) enrichment library strategy **(Figure 3.6 A)**.

Next, I evaluated the reproducibility of gene expression measurements obtained from Ribo-Seq and RNA-Seq. Between-replicate measurement error distribution was relatively narrow: on average normalised expression values from the same experimental condition were lower than 1.22 fold for 90 % of the genes **(Figure 3.6 B)**. Translation efficiency (TE), which is the ratio of normalised ribosomal footprints abundance to mRNA level, showed almost 100-fold dynamic range **(Figure 3.6 C)**, which is similar to the original Ribo-Seq protocol (Ingolia et al., 2009). Lastly, the values for both techniques were highly reproducible between biological and technical replicates, average Pearson's product of correlation was 0.965 for both Ribo-Seq and RNA-Seq **(Figure 3.6 D)**. Interestingly, when Ribo-Seq measured expression levels are juxtaposed with the RNA-Seq counterparts two observations can be made. Firstly, ribosomal footprints abundance has smaller dynamic range that mRNA abundance and, secondly, the relationship between Ribo-Seq and RNA-Seq signal seems to be intensity dependent **(Figure 3.6 D)**. This means that the high correlation between the number of ribosomal footprints and mRNA abundance can be spurious, particulairly for lowly expressed transcripts, and a translation intensity of a mRNA needs to be interpreted with care.

I conclude that all the quality measures and the reproducibility of gene expression measurements meet the expectations for Ribo-Seq data. By comparison to previous Ribo-Seq studies, this experiment has larger number of replicates (4 versus usual 2) and shows overall high reproducibility and good quality of the libraries. This confirms that ribosomal footprints obtained from primary GC B-cells: 1) correspond to the ribosomes involved in active translation, and 2) can be used for further analyses.

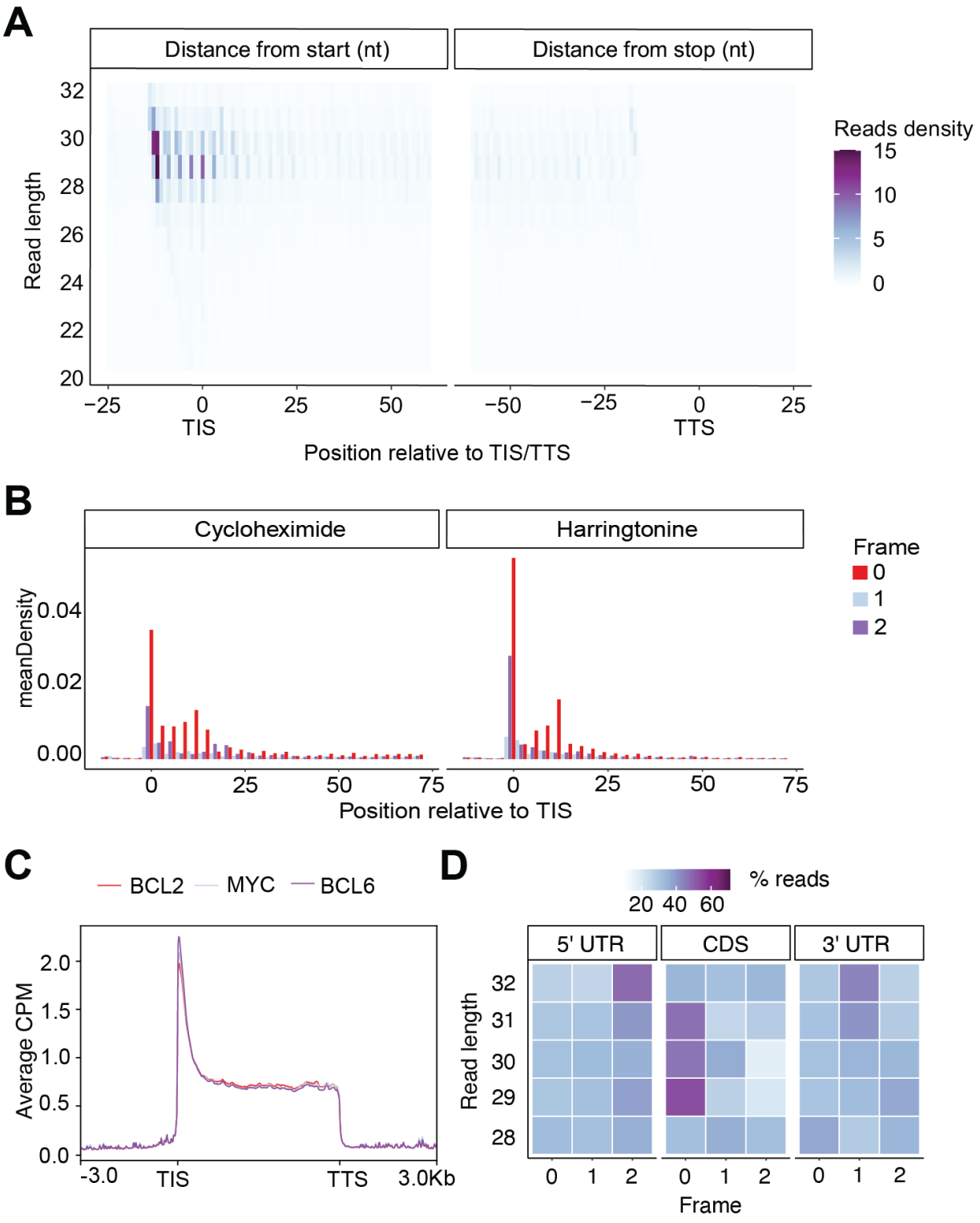

Figure 3.5: Reading frame restriction of Ribo-Seq reads for the BCL6/MYC experiment in human GC B-cells

A) Heatmap of Ribo-Seq read frame usage by read length, showing read frame restriction in the 28-31 nucleotide ribosome protected fragments.

B) Histogram of mean fold changes in TMM normalised ribosome footprints density between replicates of the same experimental condition.

C) Metagene plot showing distribution of mapped Ribo-Seq reads to regions of the transcript.

D) Heatmap of Ribo-Seq read frame usage by gene feature, showing how the characteristic read frame bias within the CDS but not the UTRs.

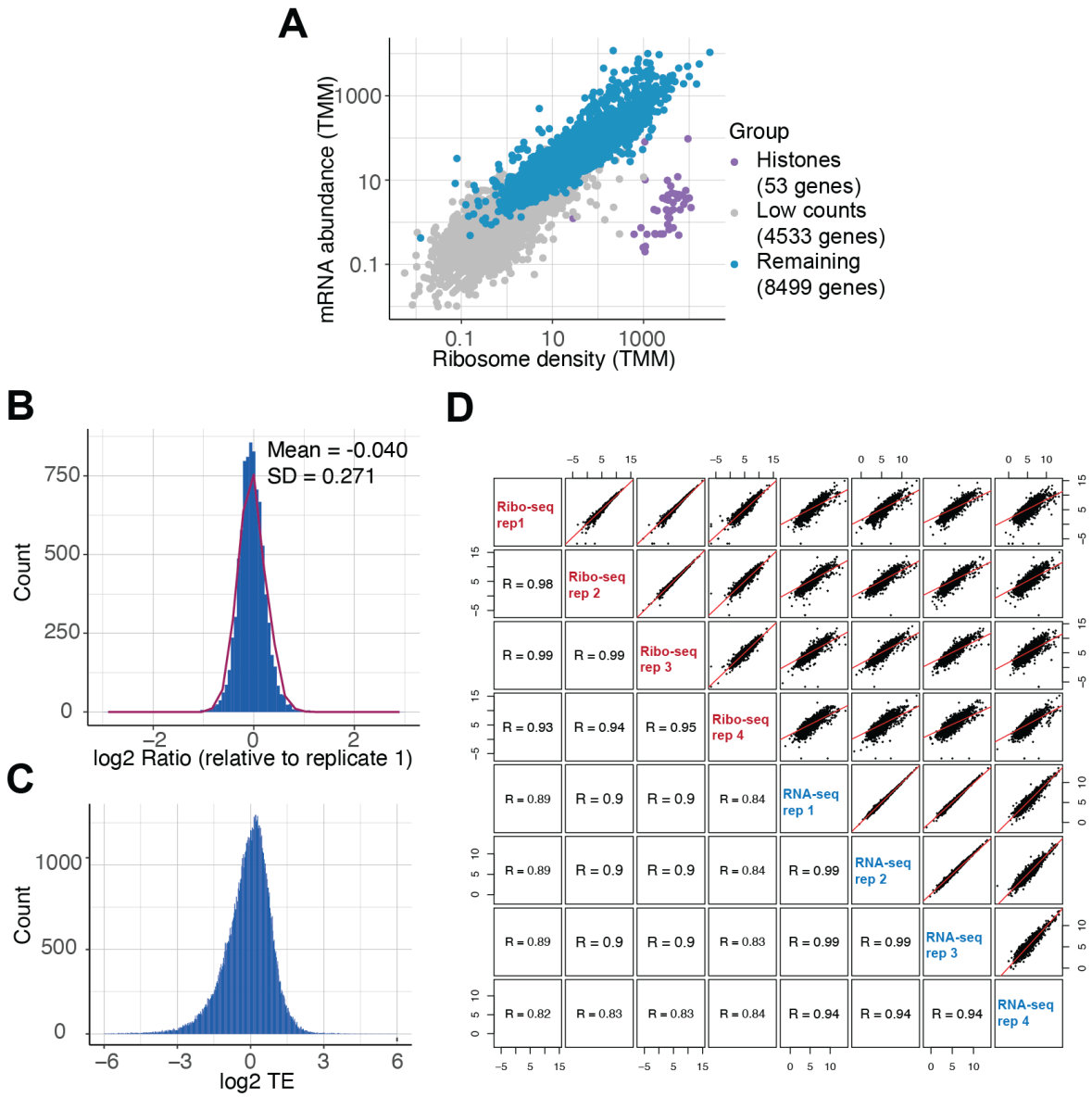

Figure 3.6: Low counts filtering and reproducibility of Ribo-Seq and RNA-Seq samples for the BCL6/MYC experiment in human GC B-cells

A) Scatter plot showing trimmed mean of M-values (TMM) normalised ribosome footprints density against mRNA abundance. Genes classified as low counts and histons were excluded from further analysis.

B) Histogram of mean log2 fold changes in ribosome footprints abundance between samples from the same experimental condition. This is similar to the reproducibility reported by Ingolia et al. (2009).

C) Histogram of mean Translation Efficiency (TE) values obtained from paired Ribo-Seq and RNA-Seq samples from the same experimental condition. Similar dynamic range was shown by Ingolia et al. (2009).

D) Scatter plots showing consistency between four replicates of Ribo-Seq and RNA-Seq. Pearson correlation coefficients are shown for each comparison.

3.1.3 Benchmarking statistical approaches for differential translation analysis

The aim of differential translation analysis is to identify genes regulated predominantly at the level of translation. Although many computational tools has been developed to perform the statistical analysis of differentially translated genes, very few have been used beyond their initial publication. The popularity of the Ribo-Seq technique, however, is growing. In 2021 the original paper (Ingolia et al., 2009) reached 3050 citation and, according to PubMed, there are almost 1000 papers containing Ribo-Seq or ribosome-profiling in the title or abstract. An opposite trend has been observed for the computational methods. Only two out of seven most popular tools for differential translation analysis, that will be reviewed in this section, exceeds 100 citation. For comparison, the number of citations for DESeq2 and edgeR, which are the two most popular packages for differential expression analysis of RNA-Seq experiment go over 27 000 and 20 000, respectively (Love et al., 2014; Robinson et al., 2010).

To make an informed decision on how to identify translationally regulated genes associated with BCL6 or MYC overexpression, I reviewed previous approaches to differential translation analysis and evaluated their performance using an external dataset of 54 Ribo-Seq/RNA-Seq pairs from Yoruba lymphoblastoid cell lines (Battle et al., 2015). The choice of this dataset has been dictated by the large number of replicates available, which allows me to benchmark tools using various experimental designs. I will return to the BCL6/MYC data at the end of this chapter.

Translation efficiency

Ribo-Seq relies on the assumption that the density of ribosomal footprints (Ribo-Seq reads) mapping to a gene is a proxy for its translation intensity (Ingolia et al., 2009). Because the more abundant mRNAs produce more ribosomal footprints, the Ribo-Seq signal is dependent on a transcript abundance. In other words, the number of ribosomal footprints mapping to a gene is positively correlated with the number of RNA-Seq reads assigned to the same gene. Therefore, each Ribo-Seq sample is always paired with its corresponding RNA-Seq sample in order to enable the calculation of mRNA-specific translation rate. Translation Efficiency (TE) is a measure introduced first by Ingolia et al. (2009) to compute translation rate relative to mRNA abundance. By definition, TE is the ratio of ribosomal footprints abundance to mRNA abundance, typically presented as log ratios. Thus TE for a gene g is:

$$TE_g = \log(\frac{RF_g}{mRNA_g}),$$

where RF and $mRNA$ stand for ribosomal footprints or mRNA abundance, respectively.

Different normalisation methods can be used to calculate these values. A simple approach is to normalise the number of ribosomal footprints per gene by the library size and the length of the CDS with Reads Per Kilobase Million (RPKM) or just by the library size, using e.g. Counts Per Million (CPM). Normalisation methods developed initially for differential expression analysis of RNA-Seq data, such as Trimmed Mean of M-values (TMM) (Robinson et al., 2010) or Relative Log Expression (RLE) (Love et al., 2014), are also applicable.

Fold change in TE has been widely adopted to identify genes with translation load that is disproportionate to the transcript abundance. However, as was pointed by Larsson et al. (2010), the TE fails to control for the effect of mRNA level fully. Bias of TE ratio can be explained mathematically by *spurious correlation*, described by Pearson in 1897 (Pearson, 1897). It refers to a situation when dependent variable (TE in this case) correlates with the mean value of the independent variable (mRNA abundance) even when mRNA and ribosomal footprints abundance are uncorrelated, which is illustrated by the following equation:

$$r_{(Y-Z),Z} = \frac{-s_Z}{\sqrt{s_Y^2 + s_Z^2}},$$

where Y is a vector of ribosomal footprints abundance, Z is a vector of paired mRNA abundance, r is the Pearson correlation coefficient and s is the sample standard deviation. Here a correlation between TE and log mRNA abundance is a function of the standard deviation of TE and mRNA abundance vector from experimental replicates. Larsson et al. (2010) conclude that this property gives rise to false positives and false negatives when TE is the only metric used to identify translationally regulated genes, especially if large shifts in mRNA abundance are expected.

Brief overview of current methods for differential translation analysis

The lack of established analytic workflow, as this is the case for RNA-Seq data analysis, suggests that the differential translation analysis is not straightforward. Despite different biological origin Ribo-Seq and RNA-Seq data share common features: both represent sequence of mRNA fragments that can be summarised as count matrices of gene counts per sample. Most Ribo-Seq specific tools take advantage of these similarities tailoring transcriptomic workflows to differential translation analysis. Comprehensive review and benchmarking of developed strategies has been missing, which complicates an informed strategy selection. In order to bridge this gap and pick an optimal strategy to analyse the data for this project, I reviewed five tools:

1. Xtail (Xiao et al., 2016),
2. Riborex (Li et al., 2017),

3. anota (Larsson et al., 2011),

4. anota2 (Oertlin et al., 2019),

5. deltaTE (Chothani et al., 2019),

in terms of their statistical approaches, performance, and the impact of experimental design on the accuracy and ability to identify differentially translated genes.

Because the majority of tools either run DESeq2 in the background or utilise similar strategies to analyse counts data, I will first outline the basics of DESeq2 analysis. In DESeq2 raw counts for each gene are modelled with negative binomial distribution (NB) using a generalized linear model (GLM). Mean values for each gene are normalised taking into account differences in the total number of reads between samples and NB parameters (mean and dispersion) are estimated from the data. With a simple experiment design (for example control and treatment), a GLM regression model will have one coefficient corresponding to the $\log_2$ fold change between experimental conditions. The significance of this effect size is typically tested with Wald statistics, as detailed by Love et al. (2014).

A straightforward implementation of the DESeq2 method for differential translation analysis is **deltaTE**, which takes a combined matrix of Ribo-Seq and RNA-Seq counts and models the difference between the two, that is specific to the experimental condition. DeltaTE models this effect introducing an interaction term to the GLM formula and test for statistical significance with Wald test. A potential problem with this approach is that using both data types to estimate NB parameters may lead to biased results.

Very similar strategy (combined analysis with interaction term) is employed by **Riborex**. However, in addition to DESeq2-based workflow, two more methods are available: edgeR-based and Voom-based. All three process Ribo-Seq and RNA-Seq reads together. It is important to mention that edgeR (Robinson et al., 2010) and DESeq2, in their principles are very similar: both use GLM methods, model counts data with NB and assume most genes are not differently expressed. Differences can be mainly attributed to processing outliers, dealing with lowly expressed genes, and normalization approach which may affect the estimation of NB parameters. Voom is based on limma, which was originally developed to analyse microarrays data (Law et al., 2014). Voom estimates the mean-variance relationship of the log-counts, generates a precision weight for each observation and enters these into the an empirical Bayes model (Law et al., 2014).

Xtail also uses NB and estimates its parameters running DESeq2 in the background separately for RNA-Seq and Ribo-Seq counts. Then, it builds joint probability matrix of two distributions: one for the difference between experimental conditions (separately for Ribo-Seq and RNA-Seq), and second, reflecting the overall difference in gene expression measured by Ribo-Seq and RNA-Seq (separately for experimental conditions). Finally,

Xtail compares the $\log_2$ fold changes of ribosomal footprints with $\log_2$ fold changes of mRNA abundance OR the difference in the disproportion in translational and mRNA response between two conditions. The final p-value is selected dependent on which of the two approaches returns more conservative results.

The last family of tools, **anota/anota2** employs a different approach. While the majority of methods focus on a simple difference in TE between two conditions, anota/anota2 rely on the analysis of partial variance (APV) and linear regression to control for cytosolic mRNA levels evaluating changes that are independent on transcripts abundance.

To sum up, all tools designed for differential translation analysis (except anota/anota2) assume read counts to follow NB, model the expression with GLM and test for the difference in TE. However, only two (Riborex and deltaTE) allows for experimental design more complex than just two conditions, e.g. including time series, sequencing batches or different biological models. Differences in the strategies are also related to the estimation of NB parameters, null hypothesis and statistical approaches to identify deferentially translated genes.

An alternative approach to identify differentially translated genes, used successfully by other groups (Sendoel et al., 2017; Hsieh et al., 2012) is to combine a standard differential expression analysis performed separately for translatome and transcriptome data with heuristic system of (arbitrary) thresholds to extract genes with biologically relevant effect sizes. An example set of rules for calling a gene differentially translated is as follows:

1. Statistically significant change in ribosomal footprints abundance (evaluated with standard DESeq2 workflow), FDR < 0.1,
2. Absolute mean $\log_2$ fold change in mRNA abundance < 0.3,
3. Absolute mean $\log_2$ fold change in TE abundance > 0.3.

This heuristic approach can be naturally extended to classify genes into distinct regulatory subtypes based on the direction and coordination between translation and mRNA levels. Genes with significant change in mRNA level, with concordant direction of translational change, are classified as homodirectional (up or down), while genes with opposite direction of change (e.g. mRNA downregulated with ribosomal footprints abundance upregulated) are considered 'buffered', see section 2.2.3. I refer to this strategy as **DESeq2T** and, for the purpose of this benchmarking analysis, focus only on the differentially translated genes.

Comparison of strategies to perform differential translation analysis

To compare the performance of existing tools in their ability to accurately identify differentially translated genes, I generated a series of Ribo-Seq analysis sets with varying number of replicates and differentially expressed genes. For that purpose I used data from a real translatome study of 54 Yoruba lymphoblastoid cell lines (Battle et al., 2015). Previous performance analyses, associated with the initial publication of the differential translation tools, used predominantly artificial data. Here, by using true experimental data we are closer to mimicking a real-life user experience. I chose to use the Battle et al. (2015) data because of the large number of paired Ribo-Seq and RNA-Seq samples of the same cell type. I downloaded raw FASTQ files from SRA and processed them using my RiboStream pipeline obtaining CDS count matrices.

First, I evaluated the performance of each tool under a NULL model. From the set of all 54 samples, I generated a series of analysis sets with varying number of replicates. As they all come from repeated sequencing of the same cell type, I do not expect any genes to be identified as differentially translated. This allowed me to estimate the frequency of type I of error (false positives). Half of the randomly selected samples in each analysis set was labelled as 'treatment', the remaining half as 'control'. No separation between 'treatment' and 'control' was observed in the PCA plot **(Figure 3.7 A)**. The tools were applied to datasets with 2-20 replicates, 50 runs per tool per replicate group in total.

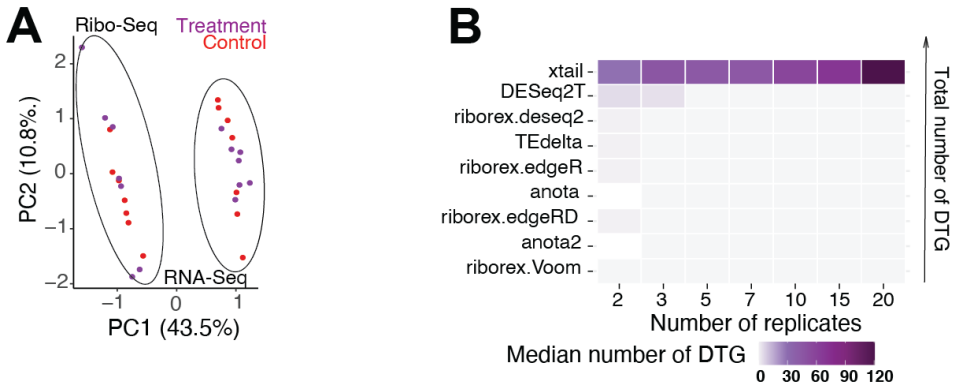

Figure 3.7: Comparison of available tools for differential translation analysis

A) Representative PCA plot showing a NULL model dataset with 7 replicates per condition. Treatment and control samples were randomly assigned from a set of 54 samples of the same cell type.

B) Heatmap showing the median number of differentially translated genes per tool per replicate group in the NULL model test (assuming no difference in expression between two conditions). For each replicate group, 50 count matrices were generated and each tool was applied.

Overall, the median number of false positive findings, at FDR < 0.05 with absolute $\log_2$ fold change in TE larger than 0.3, was the highest for Xtail (122 genes), which is in line with previous study Oertlin et al. (2019) and suggest that the performance of Xtail in terms of false positive findings may be inferior in comparison to other tools **(Figure 3.7 B)**. For other tools, the median number of false positive findings was higher in comparisons with low number of replicates. When the number of replicates was higher than 3, the number of false positives in the NULL model was negligible (with exception of Xtail).

Next, I applied the tools to a differential translation model with simulated fold changes. This time, randomly selected genes from a "treatment" group were assigned artificial fold changes sampled from a normal distribution. The fold changes were *synchronised* between RNA-Seq and Ribo-Seq samples, so that there is a 0.4 Pearson correlation between fold changes observed in RNA-Seq and Ribo-Seq, which resembles a natural relationship between mRNA and ribosomal footprints abundance **(Figure 3.8 A)**. The count matrices with designed fold changes were generated using *seqgendiff*, R package for adding a known amount of signal to a real read count matrix. Clear separation between treatment and control group was observed in PCA plot **(Figure 3.8 B)**. Similairly to the NULL model analysis, the tools were applied to datasets with 2 - 20 replicates, 50 runs per tool per replicate group (350 comparisons in total). For each tool and each run I defined a *ground truth*, which was the result of differential translation analysis performed using 20 replicates.

The first striking difference between the tools was the total number of differentially translated genes (DTGs) identified with 20 replicates **(Figure 3.8 C)**. While the TE-based strategies (TEdelta, Riborex and Xtail) identified more than 5000 DTGs, DESeq2T returned almost five times fewer genes. For anota and anota2 the mean number of DTGs was 1950 and 3600, respectively. This may reflect different definitions of differential translation between the tools. For TE-based strategies, statistically significant change in TE (which may be dependent on either change in ribosomal footprints, mRNA abundance or both) is sufficient to classify a gene as differentially translated. Whereas, for anota/anota2 and DESeq2T, the change in ribosomal footprints abundance must be significant AND independent on mRNA abundance (anota/anota2) or with negligible change at mRNA level (DESeq2T).

Next, I compared the robustness of the DTGs identification between the tools. For each tool and replicate group I computed the following metrics: false positive rate (FPR), false negative rate (FNR), true positive rate (TPR), accuracy, precision and the F1 score **(Figure 3.8 D)**. Similairly as for NULL model test, the highest rate of false positives was seen for Xtail.

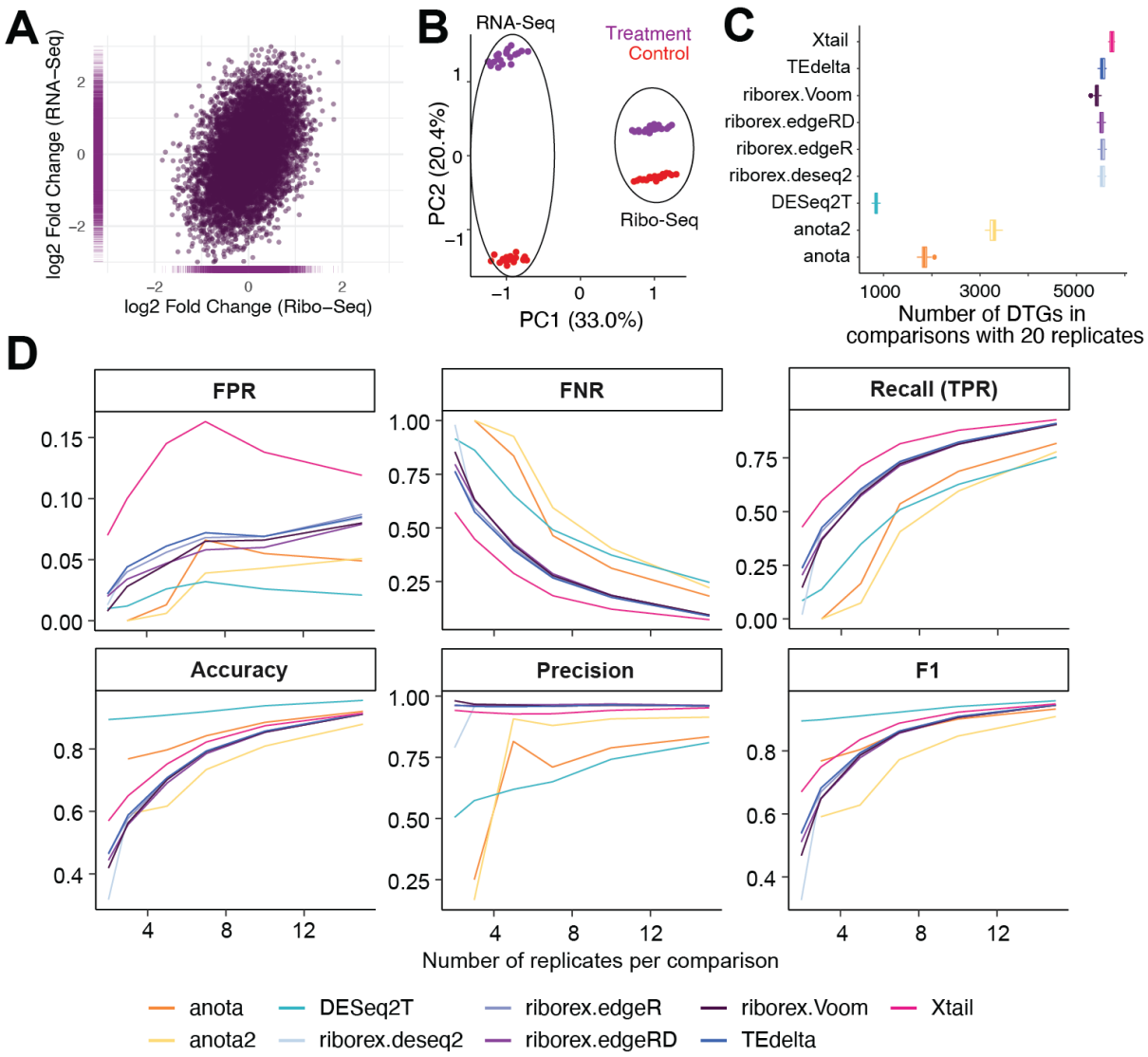

Figure 3.8: Comparison of available tools for differential translation analysis

A) Representative scatter plot of log2 fold change distribution for simulation of Ribo-Seq and RNA-Seq differential expression analysis.

B) Representative PCA plot showing a model differential translation dataset used for benchmarking.

C) Boxplot showing the total number of identified differentially translated genes (DTGs) by tool when 20 replicates were used for comparisons. Each tool has been applied to a 20 analysis sets.

D) Performance of 9 algorithms for identifying DTGs by the number of replicates used for the comparison. For each replicate group mean values of the performance metrics over 50 runs were computed.

Overall, TE-based strategies had higher TPR, lower FNR and higher precision (positive predictive value), but this was accompanied by higher FPR. In contrast, the accuracy (the total number of correct findings), and F1 score (weighed average of precision and TPR) was the highest for DESeq2T, while TE-based showed inferior performance. The number of replicates affected the performance of all methods, but with different strength. The general trend was inferior performance for lower number of replicates (2-3) with improvement for comparisons with 8-10 replicates. Interestingly, FPR was less dependent on the number of replicates: it increased steadily between 2 and 5, and, with the exception of Riborex and deltaTE, stabilised for 5 and more. For Riborex and deltaTE, FPR started growing again when the number of replicates exceeded 10.

The results of benchmarking analysis of methods for identifying differentially translated genes demonstrate that no single method is superior, but taking into account experimental design and biological question of interest, it is possible to differentiate between better and worse approaches. Firstly, a striking feature of a differential translation analysis, regardless the method, is its relatively low recall (true positive rate) and high FNR when the number of replicates is low. It seems that at least 8 replicates are needed to reach the level of 50-70 % true positive findings in the group of identified DTGs. Secondly, with FDR $<$ 0.05, the rate of false positive findings is well controlled for almost all TE-based tools, with the exception of Xtail, which shown high FPR in the NULL model and in the differential translation model. The lowest rate of false positive findings across all replicates groups was observed for DESeq2T. Lastly, the flexibility to analyse more complex experimental design, such as time series or sequencing batches, might be key for method selection. Among the benchmarked tools, this is possible in Riborex, DESeq2T and TEdelta.

I conclude that for the analysis of small to medium size translatome experiments with complex experimental design DESeq2T and Riborex are the preferential choice. However, when the primary interest of the analysis are the changes ribosomal footprints abundance that are independent of the mRNA level, the analysis performed with DESeq2T may provide better answer due to its lower rate of false positive findings. Therefore this is the approach I adopted to identify differentially translated genes in this and the next chapter.

3.2 Translational regulation following BCL6 and MYC overexpression in primary GC B-cells

Having settled on an analysis strategy to identify differentially translated genes, I returned to the experiment with primary GC B-cells. I analysed gene expression responses following MYC and BCL6 overexpression in primary GC B-cells following the DESeq2T method tested above and detailed in section 2.2.3. Briefly, I performed a standard differential expression analysis with DESeq2, separately for Ribo-Seq and RNA-Seq data. This revealed that, for both BCL6 and MYC, the fold changes in mRNA abundance and ribosomal footprint density were highly correlated (Pearson's product of correlation, 0.71 and 0.82, respectively) suggesting that mRNA-driven changes are the dominant mechanism of regulation **(Figure 3.9 A-B)**. Overall, in both experiment types, I identified 946 differentially expressed genes for BCL6 and more than four times more, 4247, for MYC, of which 31.8% and 43.7% overlapped between transcriptome and translatome analysis **(Figure 3.9 C)**.

Next, I classified differentially expressed genes into 6 regulatory groups, depending on the direction (up or down) and the type of expression response: translation only, coordinated (homodirectional) or buffered between translation and mRNA level, see section 2.2.3.

To characterise each of the potential regulatory programmes, I performed a gene ontology (GO) and pathway analysis. Top statistically significant findings are shown in **(Figure 3.9 D-E)**. In BCL6 overexpressing cells, genes showing coordinated downregulation at the level of transcription and translation were enriched for numerous terms associated with unfolded protein response and inflammation. Indeed, the expression of many core components of the ER stress pathway, including ERN1, XBP1, or DDIT3 (CHOP) was significantly reduced in both RNA-Seq and Ribo-Seq data. Pathways related to extracellular matrix organisation and signalling through receptor tyrosine kinases were enriched in the upregulated group. For MYC, I observed strong upregulation of terms related to translation and ribosome biogenesis, which is in line with previous studies on MYC regulatory network. The downregulated group consisted of terms related to inflammatory responses, G protein-coupled receptors (GPCR) and PD-L1 signalling. The latter included mainly genes related to HLA presentation. Interestingly, the decrease in mRNA abundance of CD274 (Programmed death-ligand 1, PD-L1) in MYC overexpressing cells was buffered by increase in the ribosome footprint density (TE $\log_2$ fold change = 1.280). In BCL6 overexpressing cells, a similar trend was observed - mild mRNA downregulation buffered by the ribosome abundance (TE $\log_2$ fold change = 0.42). The expression of PD-L1 on the cell surface is known to inhibit T cell-mediated immune responses, which is a well established mechanism of maintaining physiological self-tolerance, but also promoting immune escape

during tumour formation (Pardoll, 2012). Maintenance of PD-L1 translation combined with translational downregulation of HLA-related genes by MYC, could mediate permissive environment for aberrant clonal expansion of GC B-cells, thus, facilitate tumour formation. However, the average expression of PD-L1 in primary GC B-cells was low, just above 10th percentile of expression.

The mean difference in TE had narrow dynamic range. Only about 1% of genes had absolute $\log_2$ fold changes larger than 1. Translation-level changes were dominated by the regulation of synthesis of *housekeeping* proteins, such as the components of the mitochondrial respiratory chain (MYC only) or ribosomal proteins (MYC and BCL6) **(Figure 3.9 E)**. 4 out of 7 mitochondrially encoded subunits of NADH dehydrogenase (complex I) were translationally upregulated in MYC overexpressing cells (TE $\log_2$ fold change between 0.4 and 1) as well as CYCS gene (cytochrome c), and 3 components of the cytochrome c oxidase **(Figure 3.9 F)**. This may complement a known function of MYC in promoting mitochondria biogenesis (Morrish and Hockenbery, 2014).

When it comes to ribosomal proteins (RPs), BCL6 or MYC overexpression was associated with translational control of few of them, but in opposite directions **(Figure 3.9 E)**. RPL28, RPL32, RPL36AL, RPLP0 and RPS9 were translationally suppressed in MYC overexpressing cells, while translational upregulation of RPL5, RPL24, RPL18, RPS27L and RPL30 was associated with BCL6 overexpression. Translational regulation of RPs is typically associated with mTOR signalling and preferential translation of transcripts with 5′terminal oligopyrimidine (5′TOP) motifs. However, since MYC overexpression is associated with an increase in the RPs mRNA abundance, changes in TE can reflect mRNA rather than translation-driven changes.

Overall, I conclude that the overexpression of MYC or BCL6 is associated with relatively minor changes in translation intensity that are independent on mRNA-driven reprogramming. Translational control affects only selected transcripts, predominantly those involved in highly energetic processes such as ribosome biogenesis or respiratory chain. In addition, translational buffering of decreasing levels of PD-L1 mRNA were revealed, which may suggest the involvement of translational control in immune surveillance mechanisms. Although, the data presented here do not allow me to determine the significance of such adaptive mechanisms, they provide an interesting insight into the scope of changes that can accompany deregulation of MYC or BCL6 in B-cell lymphoma.

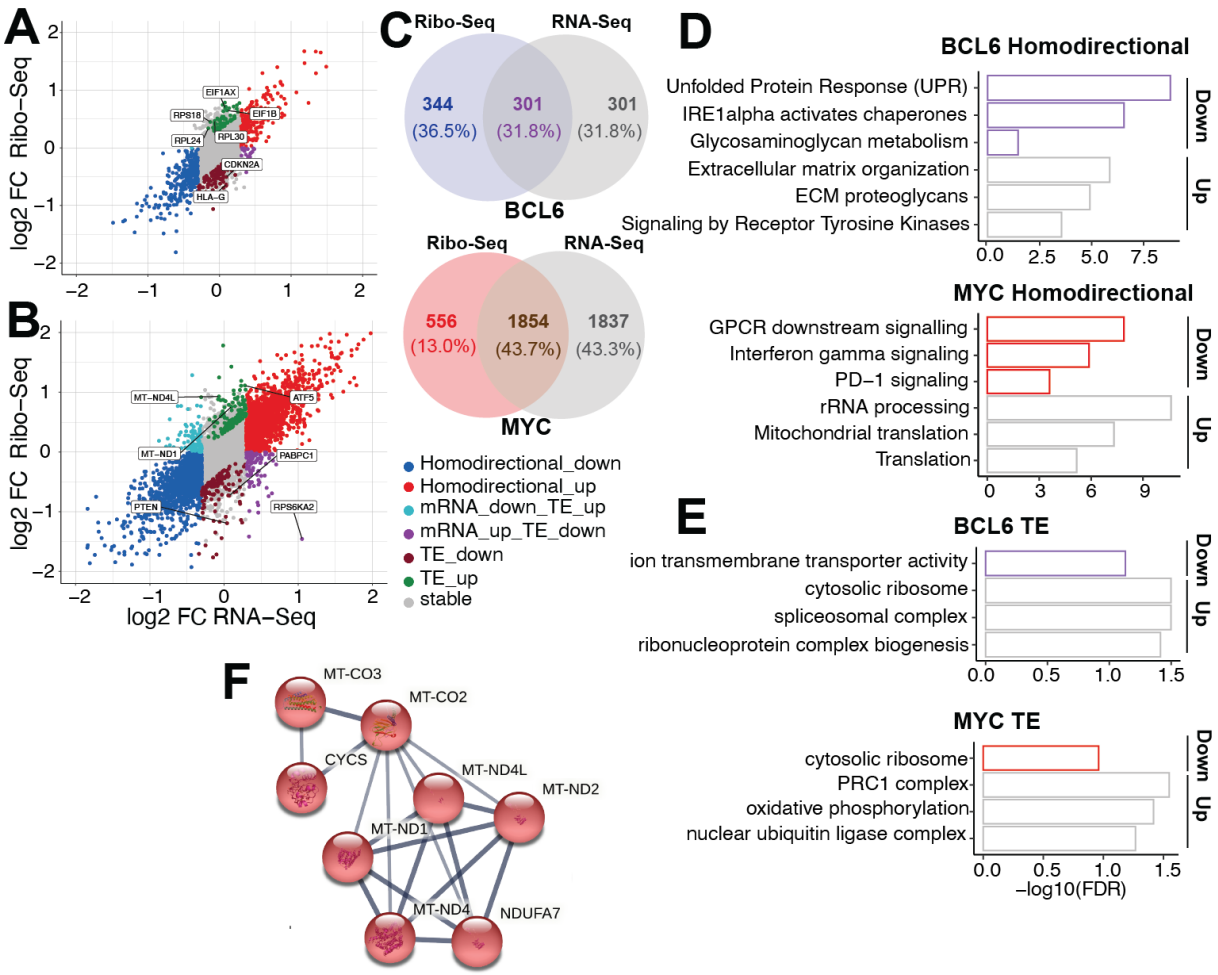

Figure 3.9: Differential translation analysis of GC B-cells overexpressing BCL6 or MYC

A) Scatter plot showing $\log_2$ fold changes in ribosome footprints occupancy against $\log_2$ fold changes in mRNA level for BCL6-t2A-BCL2 versus BCL2 comparison. Colours represent regulatory groups.

B) Scatter plot showing $\log_2$ fold changes in ribosome footprints occupancy against $\log_2$ fold changes in mRNA level for MYC-t2A-BCL2 versus BCL2 comparison. Colours represent regulatory groups.

C) Venn Diagrams showing the overlap in differentially expressed genes identified by translatome (Ribo-Seq) or transcriptome (RNA-Seq) data.

D) Reactome Pathway analysis for genes with coordinated change (homodirectional) in expression. Three top enriched pathway are shown per group (FDR < 0.1).

E) Gene ontology analysis of differentially translated genes (FDR < 0.1).

F) The network of protein-protein interactions of translationally upregulated genes following MYC overexpression. Obtained from STRING 1.1.5.

3.3 Discussion

Main findings

Ribo-seq, also known as ribosome profiling, is a Next Generation Sequencing technique that has transformed studies on mRNA translation. Here, I introduce RiboStream, a bioinformatic pipeline that I developed to process translatome data efficiently and accurately. RiboStream was applied to process a large dataset of 54 Ribo-Seq and RNA-Seq samples that allowed me to systematically benchmark available tools for differential translation analysis. Based on this, I selected DESeq2T as preferential for analysing Ribo-Seq datasets in this study. DESeq2T method combines a standard differential expression analysis using DESeq2 software with a set of rules classifying genes subject to distinct mechanisms of regulation, DESeq2T showed superior accuracy, low false positive rate and flexibility to analyse experiments with complex experimental design. A similar strategy, applied in previous studies (Hsieh et al., 2012; Sendoel et al., 2017) has lead to biologically relevant results (confirmed experimentally), despite using a lower number of replicates than in our study. I then used DESeq2T to dissect the translational consequences of BCL6 or MYC overexpression in the primary GC B-cells model. This revealed preferential translational of selected transcripts encoding certain ribosomal proteins and the components of the respiratory chain. While the concordant translational response followed changes in the mRNA level in the majority of genes, there were few exceptions from that rule. A particularly interesting example was translational buffering of CD274 expression (PD-L1).

Limitations and artefacts of translatome studies

Given the dynamic nature of translation, the process of Ribo-Seq library preparation is a critical factor in the accurate measurement of cellular translatome. In order to keep the position of translating ribosomes undisturbed during cell harvesting, the standard practice is to treat the cells with translation inhibitors just before the procedure. The most common drug used to *freeze* the ribosomes before harvesting is cycloheximide (CHX), but the fidelity and precision of ribosomal footprints mapping in samples treated with CHX has been questioned (Duncan and Mata, 2017; Gerashchenko and Gladyshev, 2014; Santos et al., 2019; Hussmann et al., 2015). Analyses performed in lower eukaryotes (*Schizosaccharomyces sp.* and *Saccharomyces sp.*) showed that CHX induces artefacts in ribosome coverage profile in cells subjected to different types of stress. CHX artefacts were responsible for skew in codon occupancy towards CGA and CGG (Hussmann et al., 2015), accumulation of the ribosomal footprints in 5′UTR and in the first 100–200 nucleotides of the coding sequence, so-called 5′ translation ramp. (Gerashchenko and Gladyshev, 2014; Tuller et al., 2010). Highly expressed genes, such as ribosomal proteins, were more prone to

experience CHX-induced artefacts (Santos et al., 2019) and the effect was dose-dependent. However, a similar study validating these findings in yeasts and mammalian cells showed that this effect is species specific (Sharma et al., 2019). Sharma et al. (2019) showed that CHX not only did not disturb the codon occupancy pattern in human cells but also did not affect the gene-level translation intensity measurements. Important factors determining the quality of the experiment included: the type of RNAse, its concentration (Sharma et al., 2019; Gerashchenko and Gladyshev, 2017) and the time between the onset of harvesting and flash-freezing (Sharma et al., 2019; Rooijers et al., 2013) These results suggest that the experiments performed in yeast cannot be simply extrapolated to mammalian systems. Although we cannot exclude that differential translation analysis performed in human cells is free from artefacts and biases, it is unlikely that the results presented here are caused by the usage of CHX during the Ribo-Seq library preparation.

A group of housekeeping genes undergo translational regulation following BCL6 or MYC overexpression

The analysis of cellular translatome and transcriptome following BCL6 or MYC overexpression revealed a profound correlation between mRNA and ribosomal footprint abundance, which suggest domination of mRNA-driven changes. Given that, during dynamic cell transitions, even up to 92 % of per mRNA-translation rates can be explained by mRNA levels (Jovanovic et al., 2015), this is not a surprising finding for proteins, which are *bona fide* transcription factors. Interestingly, the same study (Jovanovic et al., 2015), utilising a combination of pulsed-SILAC and RNA-Seq, showed that while the majority of changes in protein level is, indeed, driven by mRNA abundance, a group of *housekeeping* genes, including ribosomal proteins (RPs) and mitochondrion-related were more dependent on translation and protein degradation, respectively. The hypothesis put forward to explain this was that translational regulation is reactive to dynamic changes of cellular states by finely tuning the rate of specific metabolic processes (Jovanovic et al., 2015). Here, I identified RPs and oxidative phosphorylation related genes as translationally regulated. MYC overexpression resulted in decreased translation rate of selected RPs and increased synthesis of several key enzymes of the respiratory chain. On the contrary, BCL6 overexpression was associated with an increase in the translation of a few ribosomal proteins.

Translational control of ribosomal proteins expression

Translational regulation of RPs was reported previously. Strong repression of translation of RPs was observed during differentiation of mouse embryonic stem cells (Ingolia et al., 2011), and after treatment with mTOR inhibitor of PC3 human prostate cancer cells (Hsieh et al., 2012). RPs were also translationally upregulated in heart tissue samples of

patients with dilated cardiomyopathy compared to normal hearts (van Heesch et al., 2019). Two patterns of regulation can be distinguished in these studies: global and selective.

The first mechanism refers to 5′TOP mediated control of mTOR downstream signalling on the translation of the core components of protein synthesis machinery, which includes RPs and few translation factors (Thoreen et al., 2012; Hsieh et al., 2012; Philippe et al., 2020). While expression of MYC is known to correlate with mTOR activation (Lu et al., 2021; Pourdehnad et al., 2013; Liu et al., 2017), no such pattern has been shown for BCL6. While inhibition of mTOR signalling is associated with translational repression of all RPs (Hsieh et al., 2012), in other studies (van Heesch et al., 2019; Ingolia et al., 2011), including this one, differential translation was shown only for individual proteins. There is a possibility that forced expression of MYC or BCL6 induces adaptive mTOR-mediated adjustments of the RPs synthesis, but only a few single RPs reached the detection level. Ribosomal proteins are highly expressed; there are millions of ribosomes present in every cell, so even small changes in their abundance can have a profound effect on translation (Genuth and Barna, 2018).

Selective regulation of translation of specific RPs may be explained in the context of ribosomal heterogeneity. Tissue specific or developmental stage specific pattern of the core ribosomal proteins expression has been reported in human and other organisms, as reviewed by (Genuth and Barna, 2018; Shi and Barna, 2015). Changes in the abundance of selected RPs and the assembly of *specialised* ribosomes could drive a specific programme of translation that could facilitate the oncogenic function of MYC or BCL6. For example, knockdown of RPL28, which I found translationally downregulated in MYC overexpressing cells, is not lethal to the cell but promotes an MHC I peptides presentation of non-canonical peptides from non-AUG start codons and noncoding regions of the transcriptome. Thus, regulation of MHC I peptide presentation may facilitate immune surveillance (Wei et al., 2019). An additional level of complexity is provided by differences in the stoichiometry of RPs in free ribosomal subunits and translationally active ribosomes, or the extra-ribosomal roles of certain RPs (Shi et al., 2017). Mutations in RPs are well known in the context of ribosomopathies, hereditary disorders with increased risk of malignancy, including lymphoma. Abnormalities in the ribosome biogenesis process can trigger p53 activation and DNA damage response (Lindström et al., 2018). Therefore fine tuning of the abundance of individual ribosomal proteins may be essential for tumour initiation and maintenance.

An alternative explanation is that the disproportionate translation of specific ribosomal proteins is a marker of profound transcriptional reorganisation following BCL6 or MYC overexpression. Both Ribo-Seq and RNA-Seq analysis provides only a relative quantification of gene expression, which may mask global shifts in translation intensity. It would be interesting to investigate the changes in polysome fractions of individual transcripts and

compare them with the results from our Ribo-Seq experiment, as this could provide better resolution of translational response during dynamic reprogramming of the transcriptome.

MYC controls PD-L1 translation in primary GC B-cells

Translational control of the immune checkpoints and HLA surface molecules has attracted much attention in the context of immunosurveillance of developing Here, I observed MYC-induced downregulation of mRNA abundance of 24 out of 26 genes encoding HLA class I and II molecules combined with translational control of PD-L1, an important component of an immune checkpoint. PD-L1 is a ligand of PD1 receptor located on activated T-cells and expressed during persistent antigen stimulation to constrain immune response (Sharpe and Pauken, 2018) MYC promotes the immune escape of tumours through a variety of mechanisms. Its ability to downregulate HLA expression has been known for a long time (Versteeg et al., 1988; God et al., 2015; Staege et al., 2002), so this not comes as a surprise in MYC-overexpressing GC B-cells. MYC-induced immunosuppressive phenotype is further enhanced by transcriptional control of two immune checkpoints: PD-L1 and CD47 (Casey et al., 2016, 2018). Inactivation of *Myc* lead to abrupt decrease in PD-L1 and CD47 mRNA and protein level in multiple cancer types (*in vitro* and *in vivo* evidence) (Casey et al., 2016, 2018), which may explain an increased recruitment of immune cells to the tumour tissue following depletion of MYC (Rakhra et al., 2010). It is important to note that PD-L1 mRNA levels in cancer, including B-cell lymphoma, are also regulated by MYC-independent mechanisms, including CD274 gene amplification, translocation under active promoter or truncation of 3′UTR, which stabilised PD-L1 mRNA (Ansell et al., 2015; Green et al., 2010; Twa et al., 2014; Kataoka et al., 2016). In contrast, the overexpression of MYC in primary GC B-cells, studied here, was not associated with an increase in PD-L1 and CD47 mRNA level but almost two-fold decrease for PD-L1. PD-L1 expression level was *buffered* by an increased ribosome footprints abundance suggesting the contribution of post-transcriptional control.

PD-L1 mRNA is known to contain 2 upstream Open Reading Frames (uORF), which interrupt an effective translation of the PD-L1 protein. MYC-induced activation of the integrated stress response and subsequent phosphorylation of eIF2α was shown to alter the uORF/PD-L1 balance increasing PD-L1 expression (Xu et al., 2019). Although the ribosomal footprints abundance in 5′UTR of PD-L1 was evident in our data, there were no differences in the occupation of it between the experimental conditions. This does not allow me to exclude uORF-mediated regulation of PD-L1 translation, but it is enough to speculate that, if such exists, it is not mediated by a binary on/off translation of uORF. Interestingly, PD-L1 translation in cancer cells requires a certain *setup* of the translational apparatus. PD-L1 translation was promoted upon eIF2α phosphorylation during stress response, while the ablation of eIF4E phosphorylation at Serine 209 Xu et al. (2019) and

depletion of eIF5B reduced PD-L1 expression (Suresh et al., 2020).

So far, the PD-1/PD-L1 inhibition therapy in Non-Hodgkin Lymphoma showed limited efficiency, as most patients do not respond well to monotherapy or the response is temporary (Zhang et al., 2018b). The role of PD1/PD-L1 axis in aggressive B-cell lymphoma is complex. T follicular helper cells express high levels of PD1 and are considered important in regulating B-cell differentiation in GC B-cells and the formation of long-lived plasma cells (Goodman et al., 2017; Good-Jacobson et al., 2010). However, cell surface expression of PD-L1 was found only in 11-30% of DLBCL patients (depending on the study). PD-L1 expression was found associated with EBV positive tumours, which had inferior overall survival (Goodman et al., 2017; Kiyasu et al., 2015). A discrepancy between the direction of change upon changing MYC levels in established tumours (Casey et al., 2016), including B-cell lymphoma, and the primary GC B-cells, studied here, may suggest that the nature of PD-L1 control by MYC may change in the course of tumour development. It would be interesting to elucidate this cross-talk between the regulation of transcription, translation and mRNA stability of PD-L1, which could shed light on the susceptibility of lymphoma cells to PD-1/PD-L1 inhibition.

CHAPTER 4

Mutations in RNA helicase DDX3X facilitate MYC-driven lymphomagenesis

4.1 Background

Burkitt Lymphoma (BL) is a highly aggressive form of non-Hodgkin lymphoma (NHL) with 3:1 male:female incidence ratio (Smith et al., 2015; Morton et al., 2006). It presents in three distinct forms: endemic, sporadic and immunodeficiency-associated. The endemic Burkitt Lymphoma (eBL) is the most common childhood cancer in sub-Saharan Africa accounting for nearly half of all paediatric cancers there. The estimated number of new eBL diagnoses in Africa was 3900 in 2018 (Hämmerl et al., 2019). BL not associated with immunodeficiency and occurring outside endemic Africa is defined as sporadic Burkitt Lymphoma (sBL). It is a rare disease accounting for only 2% of all newly diagnosed NHL. The annual incidence rate of BL in Europe is about 0.36 per 100,000 (Smith et al., 2015). sBL occurs with two age peaks - 10 and 75 years.

With a growth fraction approaching 100 % and a mass doubling time around 25 hours, BL is one of the most aggressive tumours in humans. Despite such fulminant onset, it is potentially curable with intensive chemotherapy. 5-year survival is 87% for patients younger than 20 (Costa et al., 2013). The toxicity of such regimes poses a substantial risk for older individuals, thus treatment intensity must often be reduced. 5–year survival for patients 60 and older is only 25-33%. Moreover, full compliance of multi-agent treatment, which consist of sequential administration of four to six drugs, is difficult to achieve in countries with limited access to medical care. It is estimated that the long term survival of paediatric BL in sub-Saharan Africa is between 30% and 50% and has remained unchanged since 1970s (Ozuah et al., 2020).

Recent advances in molecular mechanisms of lymphoma development open new opportunities to improve the therapy of BL patients. From a molecular perspective, BL arises from the germinal centre (GC) stage of B-cell development, which is dedicated to selecting and expanding mature lymphocytes producing high-affinity antibodies (De Silva and Klein, 2015; Klein and Dalla-Favera, 2008; Basso and Dalla-Favera, 2015). This process was explained in detail in section 1.4. The oncoprotein MYC is essential for the successful outcome of the GC reaction, but its expression is transient and limited only to a small portion of GC B cells undergoing positive selection (Calado et al., 2012; Dominguez-Sola et al., 2012). MYC is a transcription factor that regulates many key cell functions, such as proliferation, DNA replication, protein biosynthesis, and metabolism. DNA-strand breaks and non-homologous end joining are the principal mechanisms involved in the immunoglobulin genes recombination. A side effect of this is the risk of translocation involving highly expressed immunoglobulin. Translocation between an oncogene MYC and immunoglobulin heavy or light chain loci is an almost universal feature of BL, observed in more than 95 % of cases (Swerdlow et al., 2016).

However, sustained MYC upregulation alone is not sufficient to drive lymphomagenesis. MYC overexpression in non-cancer cells triggers apoptosis mediated by both p53-dependent and p53-independent pathways and amplifies the apoptotic signal in mitochondria through inhibition of anti-apoptotic BCL2 expression (McMahon, 2014). A mouse model of human lymphoma revealed that further co-operating mutational mechanisms, such as PI3K activation, are needed for MYC-induced malignant transformation (Sander et al., 2012). Genes and pathways recurrently mutated in BL include transcription factors, such as TCF3/ID3, or FOXO1, SWI/SNF chromatin remodelling complex (ARID1A, SMARCA4), genes related to apoptosis (TP53, USP7, CDKN2A), GPCR signalling, B-cell receptor/PI3K signalling and epigenetic regulators (Grande et al., 2019; Schmitz et al., 2012; Bouska et al., 2017; Grande et al., 2019; López et al., 2019; Richter et al., 2012). Interestingly, in the most recent genomic study of BL (Grande et al., 2019), an RNA binding protein, DDX3X, was the most frequently mutated gene after MYC, when both point mutations and copy number changes were considered.

Despite such high recurrence rate of DDX3X mutations, little is known about their role in BL. Whilst DDX3X mutations have been also reported in chronic lymphocytic leukaemia (Ojha et al., 2015; Takahashi et al., 2018), medulloblastoma (Jones et al., 2012; Pugh et al., 2012; Robinson et al., 2012), head and neck squamous cell carcinoma (Stransky et al., 2011) and NK-T cell lymphoma (Jiang et al., 2015), the function of DDX3X in cancer remains puzzling and conflicting as it has been classified both as a tumour suppressor and an oncogene (Soto-Rifo et al., 2012; He et al., 2018). Its dual function has been reported not only for different types of cancer but also within the same type (He et al.,

2018), which underscores the context-specific effect of DDX3X in malignancy. DDX3X role in human diseases is not limited to cancer. Heterozygous mutations in DDX3X have been previously linked to neurodevelopmental disorders associated with autism spectrum, intellectual disability and seizures (Johnson-Kerner et al., 2020; Lennox et al., 2020a; Kellaris et al., 2018; Iossifov et al., 2014).

DDX3X is a highly conserved ATP-dependent RNA helicase involved in various aspects of RNA biology: transcription, splicing, nuclear export, stress granule formation and resolution, microRNA biogenesis, mRNA translation and decay (Linder and Jankowsky, 2011; Mo et al., 2021). Known RNA-independent functions include regulation of WNT, and NFκB signalling (Pugh et al., 2012; Xiang et al., 2016). DDX3**X** is located on the non-pseudoautosomal region of chromosome X and is known to escape chromosome X inactivation in wide range of tissues (Berletch et al., 2011; Cotton et al., 2015). The Y-chromosome paralogue, DDX3**Y**, shares 92% amino acid similarity with DDX3X. Although widely transcribed, DDX3Y protein is expressed exclusively in spermatogonia (Ditton et al., 2004; Rauschendorf et al., 2011; Foresta et al., 2000a).

Given the high frequency of DDX3X mutations in BL, I set out to establish the contribution of this gene to lymphomagenesis in Burkitt lymphoma.

In this chapter:

1. I examine the frequency and distribution of DDX3X mutations in BL,
2. I perform a bioinformatic analysis of multi-omic datasets to elucidate the molecular role of DDX3X in BL

This project was conducted as a collaborative project with Dr Chun Gong in the Hodson lab. Dr Gong performed the wet lab experiments whilst I performed all computational analysis.

4.2 Results

4.2.1 Examining the prevalence and distribution of DDX3X mutations

4.2.1.1 DDX3X is preferentially mutated in MYC driven lymphomas

In order to establish the frequency of point mutations in BL in the UK cohort, I reviewed the results of a 293-gene targeted sequencing panel of 39 cases of previously untreated sporadic BL. Consistently with previous reports, MYC, ID3, TP53, CCND3, DDX3X, ARID1A, FOXO1 and SMARCA4 were the most frequently mutated genes. Mutation in DDX3X was found in 30.8% (12/39) patients and was much more common in males than females, 11 versus 1, respectively **(Figure 4.1 A)**. . The same analysis was applied to 928 cases of DLBCL (Lacy et al., 2020). Of these, only 5.2% had DDX3X mutation, which was in line with other recent sequencing studies **(Figure 4.1 B)** (Chapuy et al., 2018; Reddy et al., 2017; Schmitz et al., 2012).

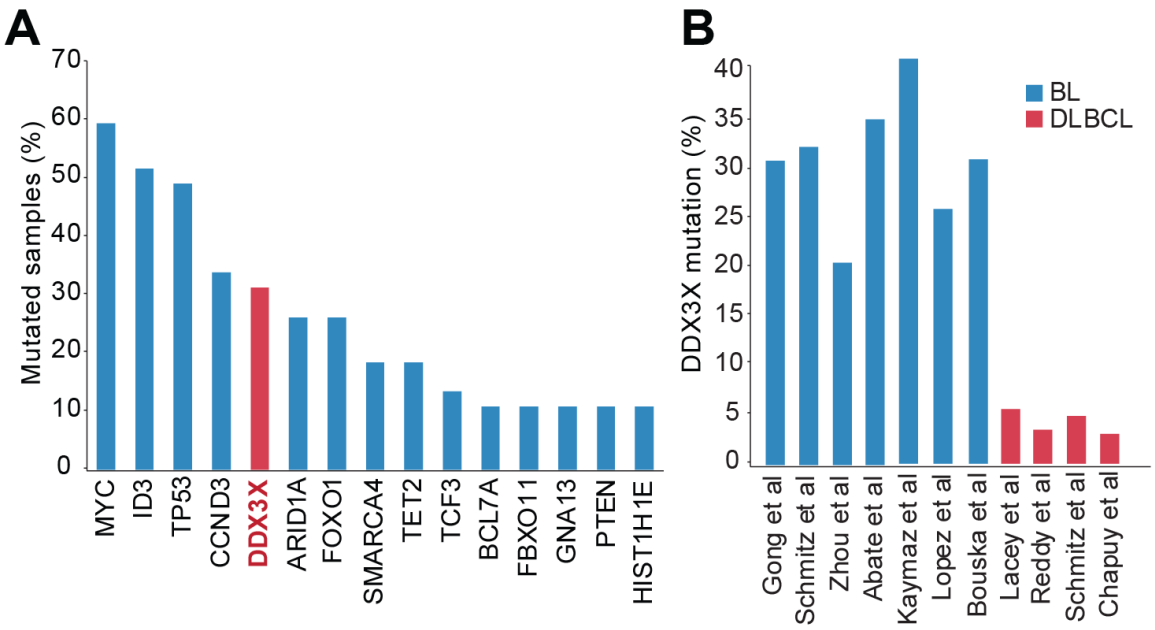

Figure 4.1: Frequency of DDX3X mutations in BL

A) Barplot showing mutation frequency (%) for the indicated genes detected using a 293-gene panel applied to 39 cases of Burkitt lymphoma (sequencing, mutation calling and filtering performed by Dr. Peter Campbell and Dr. Philip Beer).

B) Barplot showing the frequency of DDX3X mutation across published sequencing studies of BL and DLBCL.

Unlike BL, DLBCL is a highly heterogeneous disease. The recently described Molecular-High Grade (MHG) subtype of DLBCL shares several similarities with BL which are reflected by BL-like gene expression signature involving high expression of genes related to cell cycle, TCF3 signalling and ribosome biogenesis (Sha et al., 2019). In order to establish a link between DDX3X mutation and MYC-driven lymphomagenesis, I compared the relationship of DDX3X mutation frequency with MYC status in targeted sequencing data from the published UK cohort of 550 DLBCL cases with available fluorescence-in-situ hybridisation (FISH) data for MYC (Cucco et al., 2020). The frequency of MYC locus rearrangement was significantly enriched in cases with DDX3X mutation **(Figure 4.2 A**, Chi-square test, p-value = 0.001): out of 34 patients with DDX3X mutation, 16 (47.06%) showed MYC translocation, comparing with only 21.04 % (109/516) in DDX3X wild-type group. In the same study, a comparison of DDX3X mutation frequencies between DLBCL subtypes revealed remarkable enrichment of DDX3X mutation in MHG group **(Figure 4.2 B)**. Out of 558 cases with available gene expression profile, that was used to identify DLBCL transcriptomic subtype, DDX3X was mutated in 16.7% of MHG versus 4.3% and 2.2% of GCB and ABC DLBCL, respectively. The enrichment of DDX3X mutations in the MHG subtype was significantly higher than in other subtypes (p-value = 0.001, Chi-squared test).

To validate this finding, I re-analysed RNA-Seq data from a large, publicly available dataset of 553 DLBCL patients from GOYA clinical trial (McCord et al., 2019b): I downloaded the raw FASTQ files from Sequence Archive Database (SRA), performed quality control, alignment to the reference genome and read counting, as detailed in section 2.2.2. The aim of this analysis was to 1) classify the cases by their transcriptional subtypes, 2) identify those with DDX3X either mutated or not expressed and 3) compare the frequency of DDX3X alteration between the subtypes.

I combined two previously developed DLBCL classifiers to, firstly, segregate cases into ABC, GCB, Unclassified (Reddy et al., 2017) and then, distinguish the MHG group among samples belonging to GCB-DLBCL subtype, as described by Sha et al. (2019). The frequency of each identified subtype in GOYA dataset was as expected with the MHG group accounted for about 10 % of total cases **(Figure 4.2 C)**.

Next, I performed Single Nucleotide Variants (SNVs) calling from paired-end RNA-Seq data according to the GATK Best Practices Guideline. Because of the lack of germline control for each sequenced tumour, I imposed rigid criteria to establish DDX3X mutation status for each sample. I defined DDX3X mutant samples as those with either nonsense, frameshift or non-synonymous SNVs, localised in evolutionary conserved DDX3X helicase domain, with the ratio of variant coverage to reference coverage > 0.2. I filtered out all common population variants reported in the ExAC database.

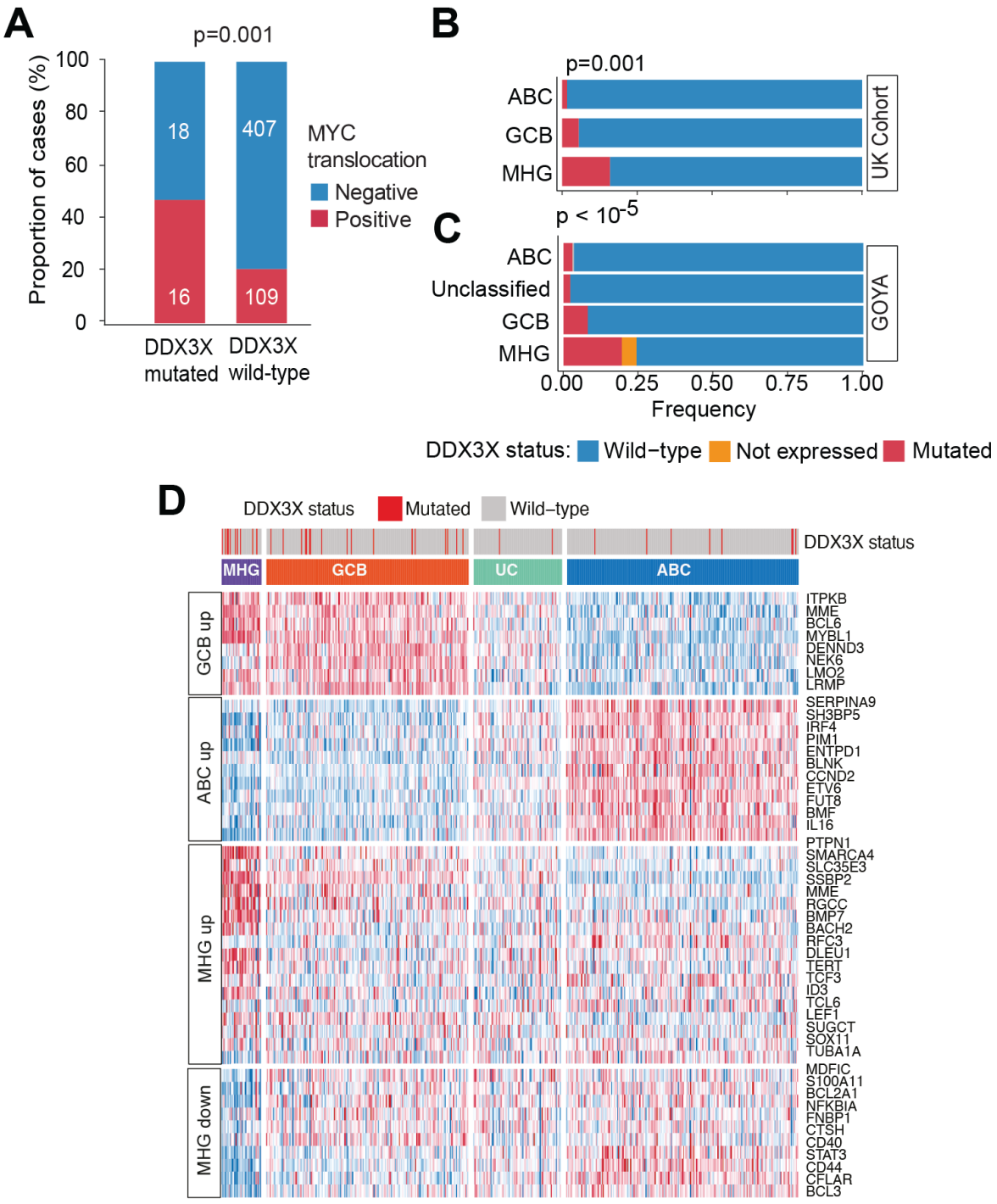

Figure 4.2: DDX3X mutations are enriched in MYC-driven DLBCL

A) Barplot showing the proportion of cases with MYC rearrangement detected by FISH from a cohort of 550 cases of DLBCL from Cucco et al. (2020) stratified by DDX3X mutation status.

B) Barplot showing the frequency of DDX3X mutation in 558 cases of DLBCL from Cucco et al. (2020) stratified by transcriptional subtype.

C) Barplot showing the frequency of DDX3X mutation in 554 cases of DLBCL from GOYA study McCord et al. (2019b) stratified by transcriptional subtype.

D) Heatmap showing DDX3X mutation status by transcriptional subtype in 553 DLBCL cases from GOYA trial. Rows represent gene expression signatures which were used to assign transcriptional subtypes (obtained from Reddy et al. (2017) and Sha et al. (2019)).

5 samples with disproportional low expression of DDX3X were also classified as *loss-of-DDX3X*. See the section 2.2.8 for a detailed description of the analytic workflow. Concordant with previous result, DDX3X mutation or absent expression was significantly enriched in the MHG subtype (19.0%), compared to GCB (8.1%) and ABC (3.0%) DLBCL (p-value $< 10^{-5}$, Chi-squared test) **(Figure 4.2 C-D)**.

The computational evidence for MYC-DDX3X interaction was subsequently tested in the lab by Dr. Chun Gong in a culture system using transduced primary human germinal centre B cells. This revealed a competitive advantage in ex vivo GC B cells when co-transduced with both MYC and helicase-mutant DDX3X. This competitive advantage was not seen in cells co-transduced with MYC and WT DDX3X or cells co-transduced with mutant DDX3X and BCL6. These functional experiments confirmed the computational prediction of co-operation between MYC and helicase-mutant DDX3X.

4.2.1.2 Context dependent pattern of DDX3X mutation in different cancer types

Since DDX3X was reported to be both a tumour suppressor and an oncogene, I wondered if there are any differences in the types of point mutations between cancer types suggesting the differential role of DDX3X. Hence, I examined the characteristics of DDX3X mutations downloaded from COSMIC, a Catalogue of Somatic Mutations in Cancer. Although the total number of mutations reported in COSMIC is biased by the availability of sequencing datasets corresponding to each tissue type, an interesting observation can be made by looking at the ratio of disrupting mutations such as nonsense or frameshift. While the majority of cancers had a varying mixture of disrupting and non-disrupting mutations, almost all mutations reported in Central Nervous System (CNS), thyroid and pancreas were of missense type **(Figure 4.3 A)** This supports the observation that the role of DDX3X mutations is context-dependent and may differ between cancer types.

4.2.1.3 DDX3X mutations in B-cell lymphomas cluster within C-terminal helicase domain

As a member of the DEAD-box RNA helicase family, DDX3X (and DDX3Y) contains an evolutionarily conserved helicase core of two RecA-like domains (Linder and Jankowsky, 2011; Mo et al., 2021). The two core domains contain 12 sequence motifs that are involved in either RNA binding (motifs Q, I, II/DEAD, VI) or ATP binding and hydrolysis (Ia, Ib, Ic, IV, IVa, V, VI). The helicase core is surrounded by two Low Complexity Domains (LCDs) that are known to participate in the assembly of RNA-protein aggregations, such as stress granules, through liquid-phase separation (Molliex et al., 2015; Valentin-Vega et al., 2016). I investigated the distribution of DDX3X point mutations across the protein domains in patients diagnosed with BL or DLBCL.

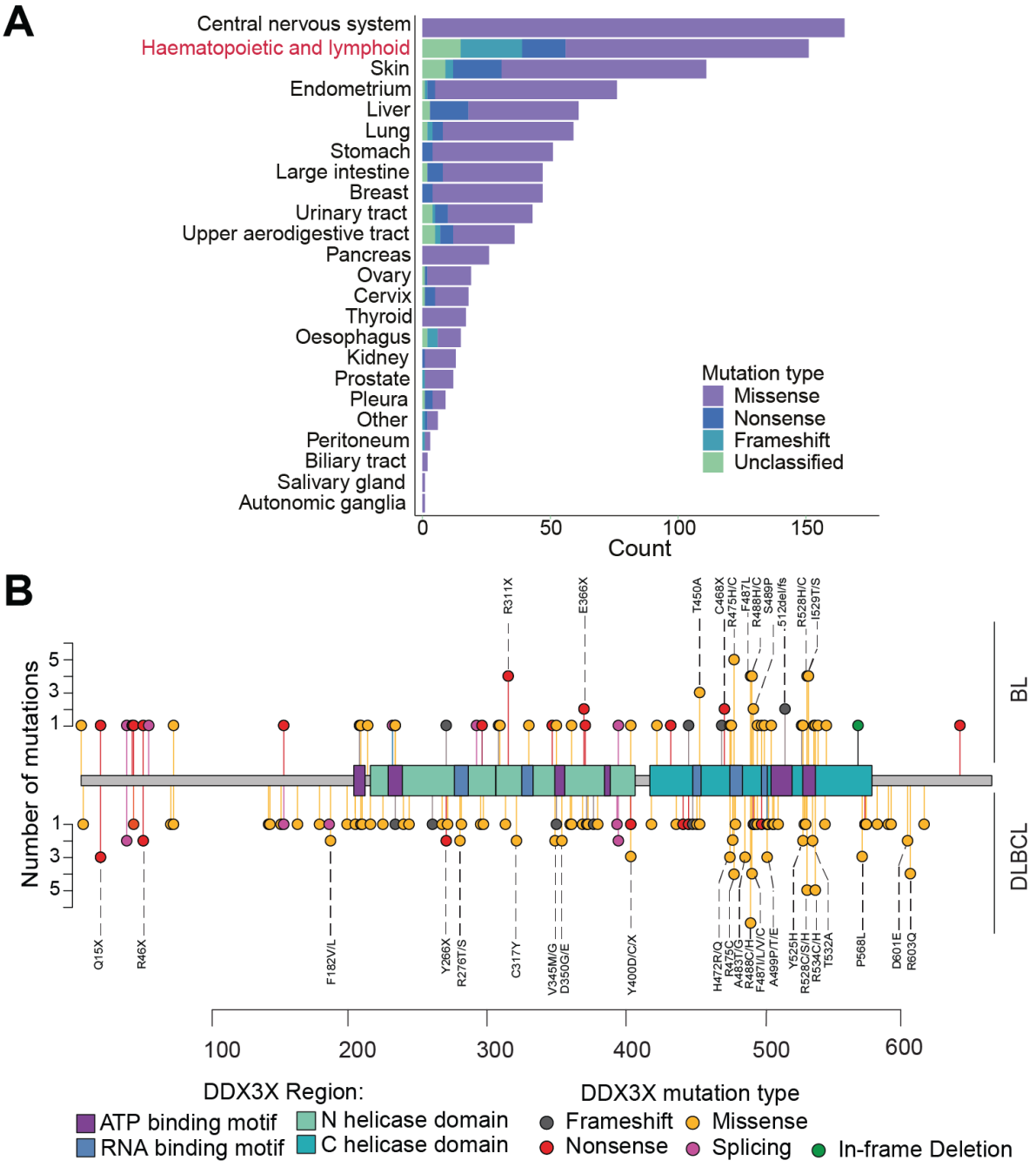

Figure 4.3: Types of DDX3X mutations in different cancer types

A) Barplot showing distribution of DDX3X mutation types in different cancer types included in COSMIC database (v.89)

B) Lollipop plot with the distribution of DDX3X mutations identified in this and published studies of BL and DLBCL over the functional domains and motifs of DDX3X protein.

This revealed the accumulation of mutations within the helicase domains, especially in the C-terminal helicase domain. The mutational hot-spots in BL and DLBCL were: R488, R475, R311, and R528, R534 **(Figure 4.3 B)**. Some of them are shared with medulloblastoma and are known to abolish helicase activity. In previous studies, RNA unwinding assays showed complete loss of helicase activity for R475 mutation (Lennox et al., 2020a) and almost 100-fold reduction in activity for R534 (Floor et al., 2016). In line with the mutation pattern observed in COSMIC data and frequent deletions of DDX3X loci reported in the recent whole-genome sequencing study of BL (Grande et al., 2019), there were multiple disrupting mutations, many of them localised near N-terminus.

These findings suggest that loss of RNA helicase function is the predominant consequence of DDX3X mutation in B-cell lymphoma.

4.2.1.4 Males with Burkitt Lymphoma and DLBCL are more likely to have DDX3X mutation

Given the strong male skew in DDX3X mutation frequency observed in the targeted sequencing panel of BL patients, I attempted to establish whether there is any relationship between sex and the probability of DDX3X mutation in B-cell lymphoma. In order to and improve confidence and precision of the effect size estimation, I performed meta-analysis of the sex skew in DDX3X mutation ratio using previously published sequencing studies of BL and DLBCL with available sex data. I collected data from 7 published studies with available DDX3X mutation status and patient sex. Data from the targeted sequencing of 39 patients from this study were also included.

Because patient sex was unavailable for 553 DLBCL patients from GOYA clinical trial, I decided to classify RNA-Seq samples into male and female using the decision tree algorithm implemented in *rpart* R package. In total, the number of patients included in the analysis was 395 for BL (6 studies) and 2180 for DLBCL (3 studies). Pooled male to female ratios were 4 for BL and 1.2 for DLBCL. The median percentage of DDX3X mutated samples was 30.385% for BL and 5.967 % for DLBCL. Single data points and sequencing type used to call variants shown in **(Figure 4.4 A - B, Table 4.1)**.

The analysis revealed that males diagnosed with BL or DLBCL have approximately 1.23 times the risk of DDX3X mutation compared to females (**(Figure 4.4 A-B)**, Random effects model p-value = 0.0002).

The allele frequencies suggested that DDX3X mutations were predominantly clonal in both females and males. In females only one copy of DDX3X was mutated **(Figure 4.4 C)**.

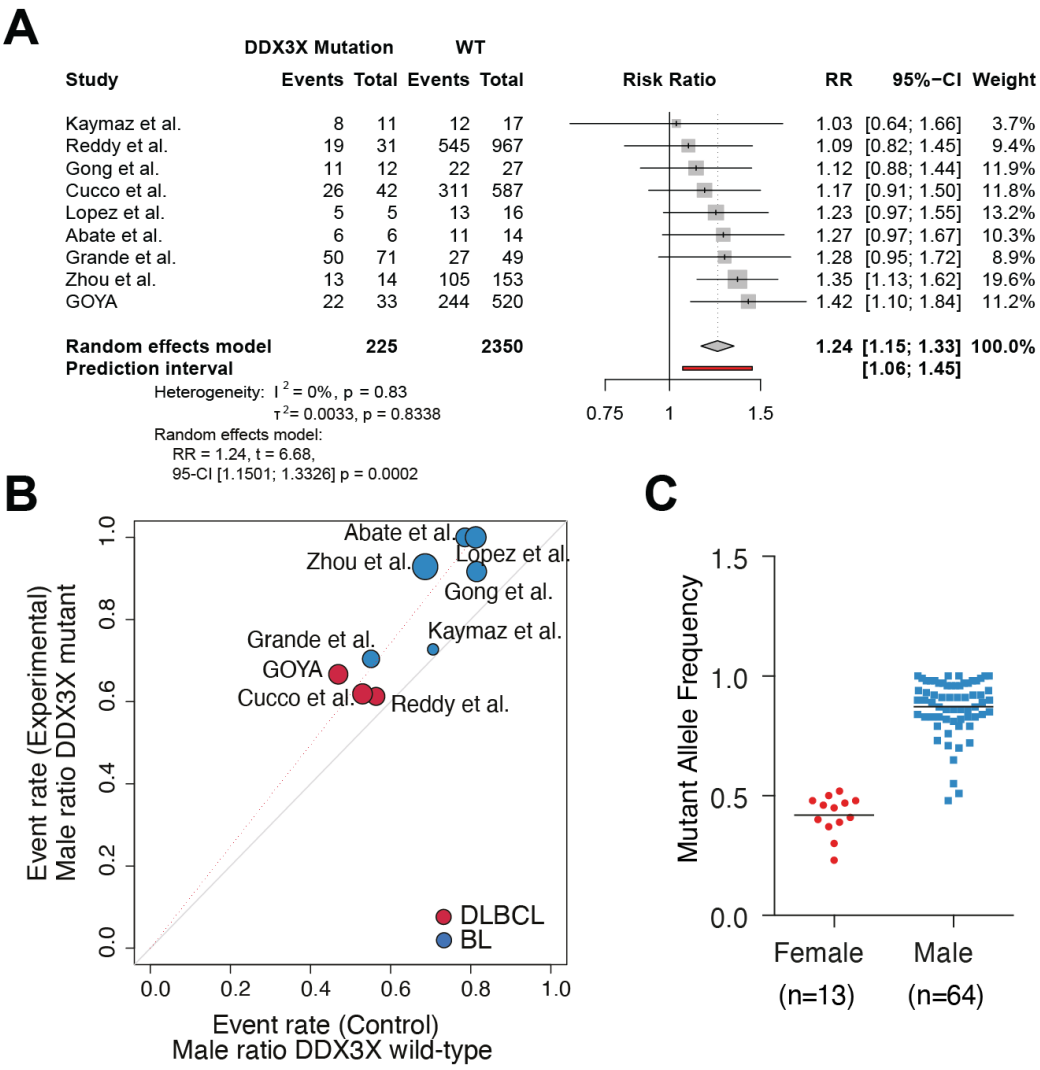

Figure 4.4: Metaanalysis of DDX3X mutation associated gender skew in published DLBCL and Burkitt Lymphoma studies

A) Forrest plot showing the effect sizes for each study. The red bar represents the prediction interval around the pooled effect shown by the grey diamond.

B) L'Abbé plot showing sex skew of DDX3X across this and other studies of BL and DLBCL.

C) Dot plot showing DDX3X mutant allele frequency by sex. Data are taken from this study and three other BL sequencing studies for which sex and MAF was available (Grande et al., 2019; López et al., 2019; Zhou et al., 2019).

Mantel-Haenszel random-effects model was used to calculate the overall Relative Risk (RR) and 95% CI. The RR in all studies had a range of 1.06–1.45 and RR of 1.23. The heterogeneity across studies was assessed by Cochran's Q test and Tau-squared (p-value 0.8338, $^2 = 0.0033$). All computations were performed using meta R package.

Table 4.1: An Overview of BL and DLBCL datasets used for meta-analysis of sex skew of DDX3X mutation occurrence

Study	Disease	Number of patients	Male:Female ratio	Sequencing type
Abate et al. (2015)	BL	20	5.67	RNA-Seq (2x75 bp)
Gong, Krupka et al. (2021)	BL	39	5.50	Targeted panel)
Kaymaz et al. (2017)	BL	28	2.50	RNA-Seq (2x100 bp)
López et al. (2019)	BL	21	6.00	WGS
Zhou et al. (2019)	167	20	2.41	WES
Grande et al. (2019)	BL	120	1.79	WGS
Cucco et al. (2020)	DLBCL	337	1.15	Targeted panel
McCord et al. (2019b) (GOYA)	DLBCL	553	0.93	RNA-Seq (2x50 bp)
Reddy et al. (2017)	DLBCL	998	1.30	WES

4.2.2 DDX3X regulates ribosome biogenesis and global protein synthesis

4.2.2.1 DDX3X binds preferentially to mRNA encoding components of core translation machinery

The existing literature suggests that DDX3X has a versatile role in cell biology encompassing many aspects of RNA biology, regulation of cell proliferation, stress response and apoptosis. In order to uncover which of these may be relevant in lymphoma, Dr. Chun Gong performed immunoprecipitation of the endogenous DDX3X and SILAC mass spectrometry of the interacting proteins (Figure 4.4 A). Gene ontology analysis of proteins interacting with DDX3X in at least one cell line revealed a strong enrichment for the proteins participating in the translation initiation including almost all components of the eIF3 complex, eIF4A, eIF4E, and eIFG4 (Figure 4.4 B). This is in line with previous works (Lee et al., 2008a; Soto-Rifo et al., 2012; Shih et al., 2008) reporting association of DDX3X protein with translation initiation complex. Among interacting proteins, there were also 7 components of stress granules (SGs). SGs are membraneless assemblies of messenger ribonucleoproteins (mRNPs) that form from mRNAs stalled in translation initiation in response to stress (Protter and Parker, 2016; Buchan and Parker, 2009; Jain et al., 2016). SGs-associated proteins included: DDX1, ATXN2L, NUFIP2, PDCD4, USP10, UPF1, and EWSR1.

These findings suggest that the role of DDX3X in lymphoma cell lines focuses on translation and protein synthesis, either through participation in translation initiation or stress granules assembly.

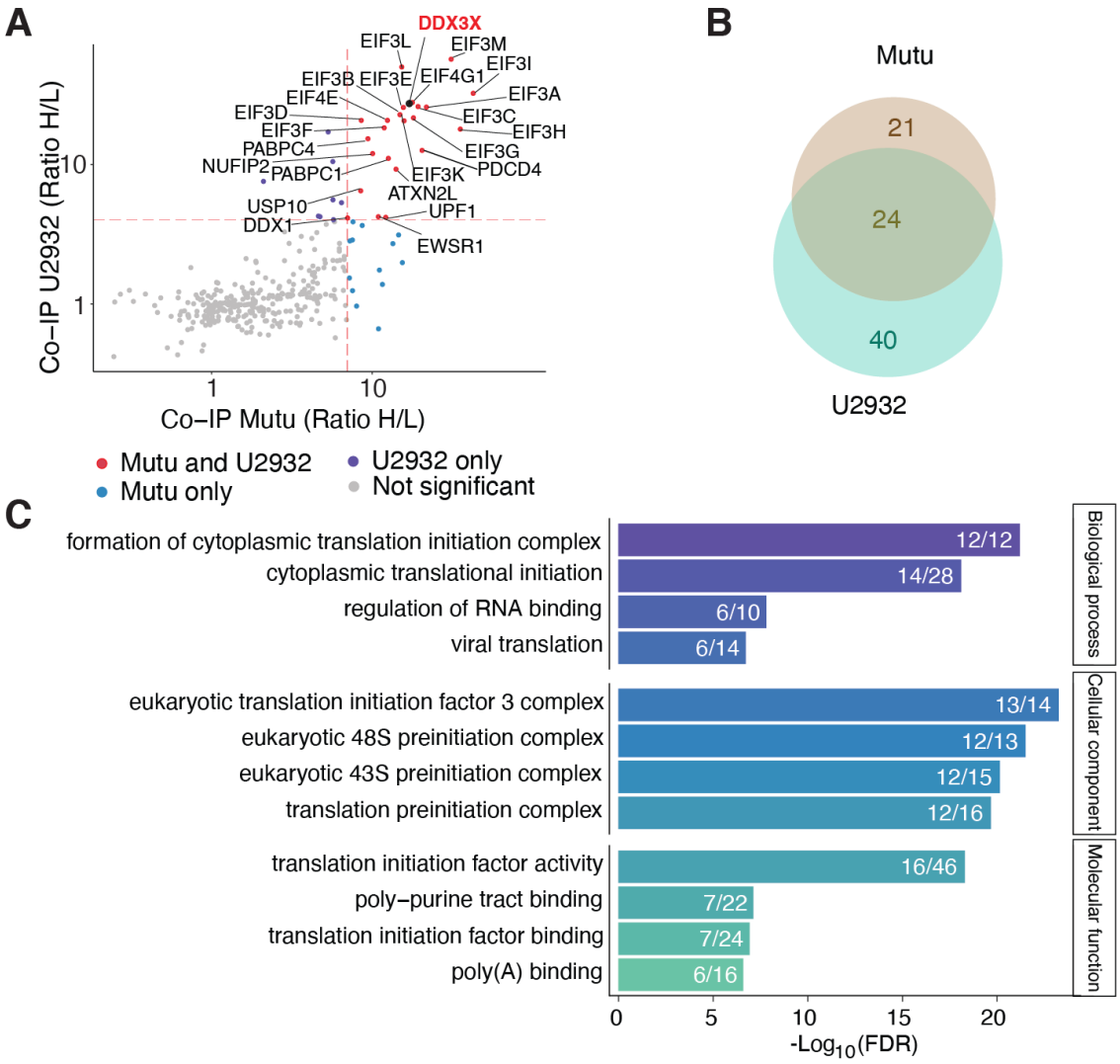

Figure 4.5: DDX3X co-immunoprecipitates with essential components of translation machinery

A) DDX3X-interacting proteins were identified by SILAC-MS following immunoprecipitation of endogenous DDX3X in U2932 (RRID: CVCL_1896) and Mutu (RRID: CVCL_ZY05). Scatter plot shows log_2 SILAC ratios of interacting proteins. Proteins significantly enriched in both cell lines are labelled. The experiment was performed by Jade Gong.

B) Venn diagram showing overlap of DDX3X-interacting proteins in two lymphoma cell lines - Mutu and U2932

C) Barplot showing gene Ontology (GO) enrichment of DDX3X-interacting proteins identified in both cell lines.

As an RNA-helicase, DDX3X can bind directly to RNA affecting its fate and function. In order to identify transcripts bound by DDX3X, we decided to perform individual nucleotide resolution crosslinking immunoprecipitation (iCLIP). This technique combines immunoprecipitation of UV-crosslinked protein-RNA complexes with Next Generation Sequencing, which allows mapping the localisation of protein-RNA complexes with single nucleotide precision (Hafner et al., 2021). Although a similar technique was used previously by other groups, they all used HEK293T cells transfected to expressed FLAG-tagged DDX3X (Valentin-Vega et al., 2016; Oh et al., 2016; Calviello et al., 2021). To investigate the binding profile of DDX3X at physiological expression levels and take into account the context-specific functions of DDX3X, it was crucial to perform the experiment with endogenous DDX3X protein in lymphoid cells. iCLIP was conducted using two lymphoma cell lines: U2932 and Mutu, as well as non-malignant human GC B cells purified from discarded tonsil tissue (Caeser et al., 2019). The experiment was performed by Dr. Chun Gong in at least two biological replicates per condition with an isotype control with IgG antibody that does not recognise DDX3X. Details regarding the experimental protocol are described in (Gong, Krupka et al., 2021).

Overall, the number of uniquely mapped reads for DDX3X iCLIP was between 4-24 million which, as expected from this technique, accounted for about 30-40% of the total number of reads aligned. In contrast, the number of uniquely mapped reads in the IgG control samples was less than 15,800 suggesting high signal-to-noise ratio. Because the density of crosslinking sites per gene was highly consistent between replicates (Pearson's Correlation Coefficient between 0.7 - 0.96), I pooled the data from the same cell type together to increase the sensitivity of DDX3X binding analysis, see methods 2.2.7.1.

First, I examined the distribution of DDX3X crosslinking sites across genomic regions. In all cell types DDX3X bound ubiquitously to mature coding transcripts **(Figure 4.6 A)**.

To visualise the precise location of the DDX3X binding site from a mature mRNA perspective, I performed a metagene analysis where the aggregated coverage of all crosslinking sites over all expressed transcripts is plotted relative to the known translation start and termination site. This showed strong enrichment in DDX3X binding at translation initiation sites (TIS) and further into the open reading frame with another peak in crosslinking sites density at the translation termination site (TTS) **(Figure 4.6 B)**. These findings are consistent with our previously demonstrated association of DDX3X with the proteins of translation initiation machinery **(Figure 4.5)**.

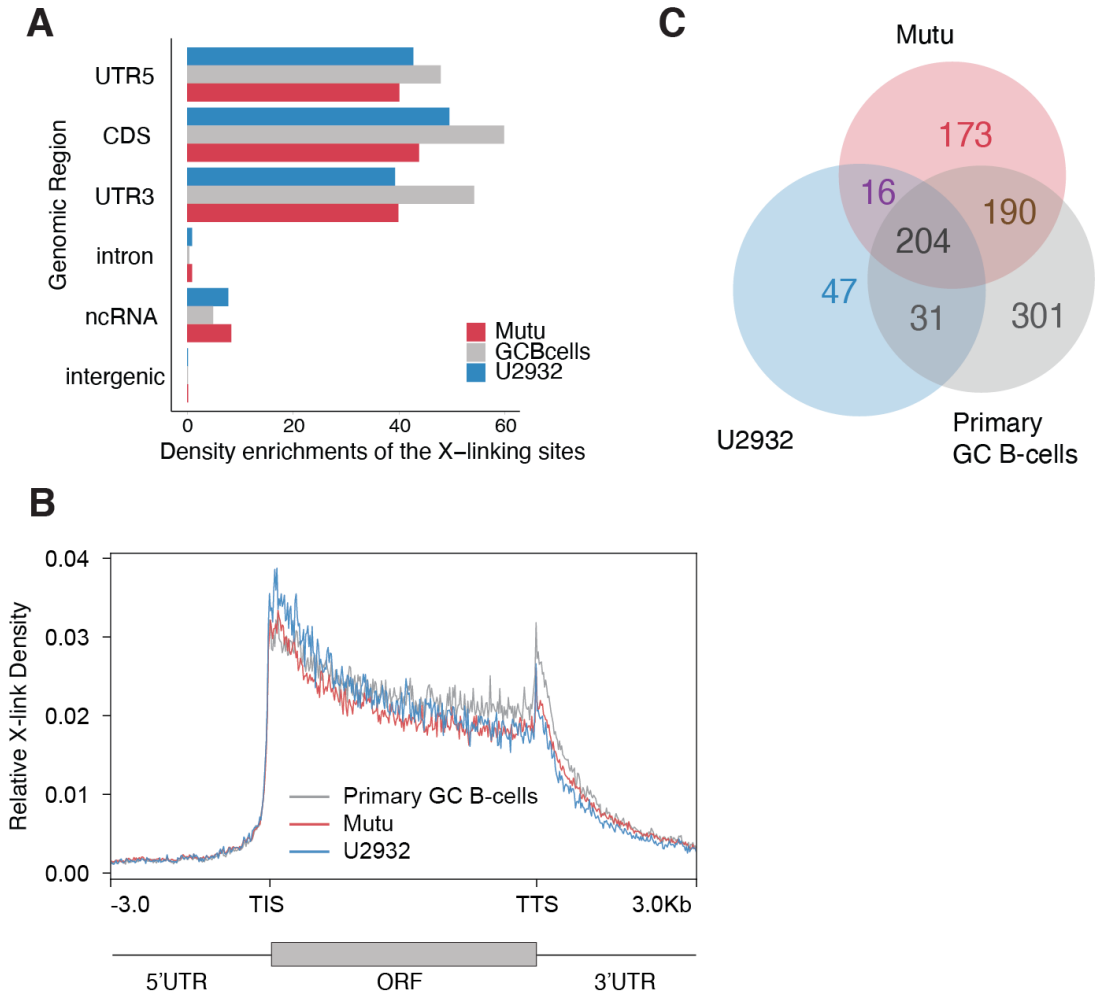

Figure 4.6: Binding profile of endogenous DDX3X in lymphoid cells

A) Barplot showing density of iCLIP cross-link sites mapping to the indicated genetic features is shown for the indicated cell types.

B) Venn diagram showing the overlap between DDX3X-bound transcripts detected in iCLIP experiments in lymphoma cell lines and primary human GC B cells.

C) Metagene summary of cross-link density across DDX3X-bound mRNA transcripts, showing length-scaled coding region and 3kb of the 5' and 3' untranslated regions. TIS = translation initiation site, TTS = translation termination site, ORF = open reading frame.

As an RNA-helicase DDX3X is known to facilitate translation of transcripts with complex secondary structures in their 5′UTR. To address this behaviour I investigated the preference in DDX3X binding with regards to the adjacent sequence context, GC content and RNA secondary structure. *De novo* motif search revealed no consensus binding motif, neither the analysis of folding energy profile and GC content around the binding site showed no pattern. This suggests that the binding of DDX3X to mRNA is not directly related to any particular RNA sequence context.

Next, I examined whether DDX3X binds to a specific family of transcripts. In order to define a list of high-confidence DDX3X targets, I filtered out first all binding peaks with less than 10 crosslinking sites and then, discarded genes with less than 3 iCLIP peaks. I found substantial overlap between DDX3X targets in all three cell types - 45.73 % of all high-confidence targets were detected in at least 2 cell types **(Figure 4.6 C)**. The majority of cell type-specific targets could be explained by the difference in the expression level of the transcript and by the difference in the total number of uniquely mapped reads between the groups. Although I had previously hypothesised that the R475 mutation situated in the RNA-binding domain may abolish interaction with RNA, the strong overlap of iCLIP targets between Mutu (DDX3X R475S) and the WT cells (U2932, primary GC B-cells) is evidence that this is not the case. However, I acknowledge that subtle binding differences are not excluded.

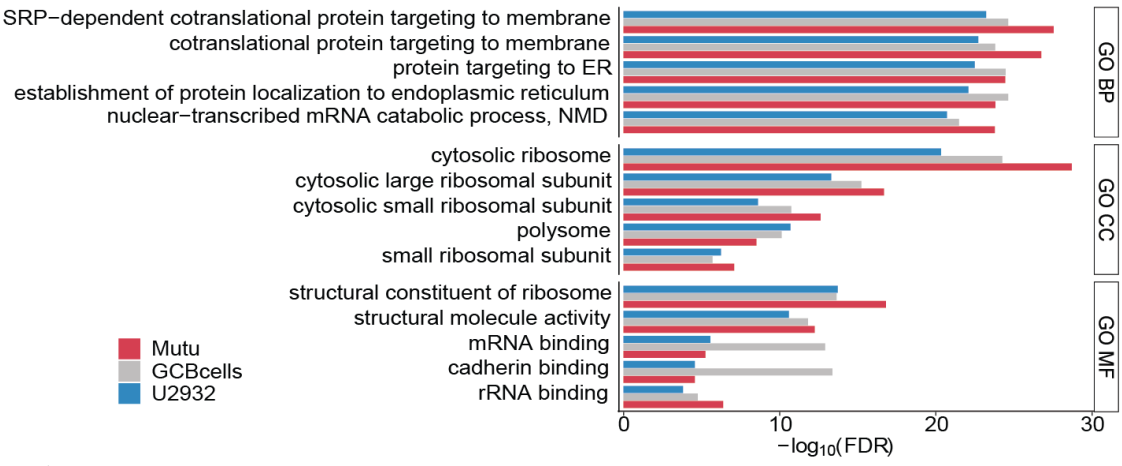

Figure 4.7: DDX3X protein binds predominantly to mRNA of ribosomal proteins
Barplot showing Gene Ontology (GO) enrichment of DDX3X-bound transcripts identified by iCLIP in the indicated cell types.
BP = Biological Process, CC = Cell component, MF = Molecular Function.

Gene ontology analysis of mRNA targets that were shared between at least two cell types (441 genes) revealed a strong enrichment for mRNAs encoding components of the core translation machinery, in particular, 45 ribosomal proteins, 9 translation initiation factors, 67 genes associated with cellular response to stress and 23 3′UTR binding genes **(Figure 4.7)**. High expression level of the core components of translation machinery enriched here may be considered a confounding factor suggesting that the iCLIP signal is non-specific. However, the iCLIP peaks spans the broad range of expression levels and these highly expressed mRNAs do not come up in other iCLIP experiments using identical protocol (personal communication with Dr. Martin Turner), which supports the specificity of the DDX3X binding.

These findings suggest that DDX3X binds preferentially to mature transcripts encoding proteins linked to various aspects of protein translation.

4.2.2.2 DDX3X regulates translation of a subset of expressed transcripts

Links between DDX3X and protein synthesis machinery were revealed by: 1) co- immunoprecipitation of DDX3X with several translation initiation factors and stress granules components, and 2) the observation of preferential binding to a subset of mature transcripts, prompted me to form a hypothesis, that DDX3X may alter the translation of its mRNA targets. To determine which transcripts are sensitive to DDX3X depletion, we performed transcriptome-wide translational profiling (Ribo-Seq) in lymphoma cell lines. The basic assumptions, strengths and limitations of Ribo-Seq technique were discussed in detail in Chapter 3. Sequencing libraries were prepared by Dr. Jie Gao as detailed in section 2.2.2.2 . The experiment was performed in two lymphoma cell lines: Mutu and U2932 with two different DDX3X shRNA and a scrambled shRNA as control. For each sample, a paired sequencing (RNA and Ribo-Seq) was performed at two time points (24h and 48h) **(Figure 4.8)**.

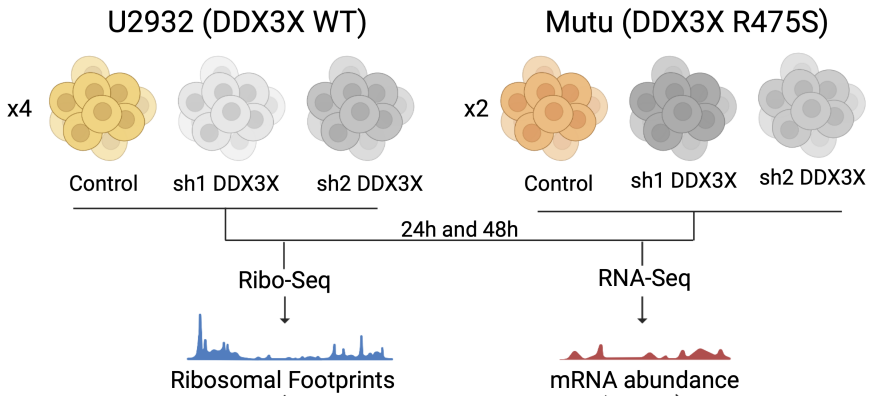

Figure 4.8: Identification of DDX3X sensitive transcripts: Experimental setting
Diagram showing experimental design of translational profiling for DDX3X shRNA cells

The average number of uniquely mapped reads (non-rRNA) in the Ribo-Seq samples was 9,623,015. In all samples, the profile of aligned sequencing reads meets expectations for Ribo-Seq experiment:

1. fragment length between 26 and 32 nucleotides with a peak at 28, nucleotides,
2. enrichment of fragments in CDS,
3. evidence of three-nucleotide periodicity in the frame preference **(Figure 4.9 A-C)**,
4. characteristic pattern shown in the metagene analysis of the estimated P-site location showing enrichment of footprints at the start codon and abrupt drop-off at the stop codon **(Figure 4.9 D)**, performed as described in section 3.1.1

The gene expression measurement with Ribo-Seq and RNA-Seq were highly reproducible between experimental replicates. The Pearson correlation coefficient was higher than 0.97 for all the samples from the same sequencing type. mRNA levels were also strongly correlated with the ribosome abundance (Pearson correlation coefficient 0.92-0.95). All this together supports a good quality of the experiment.

I sought to classify transcripts into distinct regulatory profiles. The extent to which change in expression occurs at the level of translation, transcription or both can be distinguished by juxtaposing the difference in ribosomal footprints with the difference in mRNA abundance. Because mRNA abundance positively correlates with the number of ribosomal footprints mapping to this region, the ribosomal footprints density for each transcript should be normalised by the transcript abundance. The obtained metric reflects the relative density of ribosomes per transcripts and is known as Translation Efficiency (TE). Interpretation and limitation of TE has been reviewed in section 3.1.3. I dissected mRNA and translation driven changes using the method described in section 3.1.3. This is the same strategy that was applied to elucidate translational consequences of BCL6 and MYC overexpression in the chapter 3.

This analysis revealed that the changes in expression profile after DDX3X depletion were mainly limited to the level of translation. In DDX3X WT cell line U2932, out of 200 differentially expressed genes, 70 and 90 showed decreased or increased translation rates respectively (**Figure 4.10 A**). Because the number of genes differentially expressed at the level of mRNA was relatively small (40 genes), I decided to simplify the original classification described in section 2.2.3. The remaining 40 genes differentially expressed in RNA-Seq data were classified into two groups: mRNA down or mRNA up, depending on the direction of change.

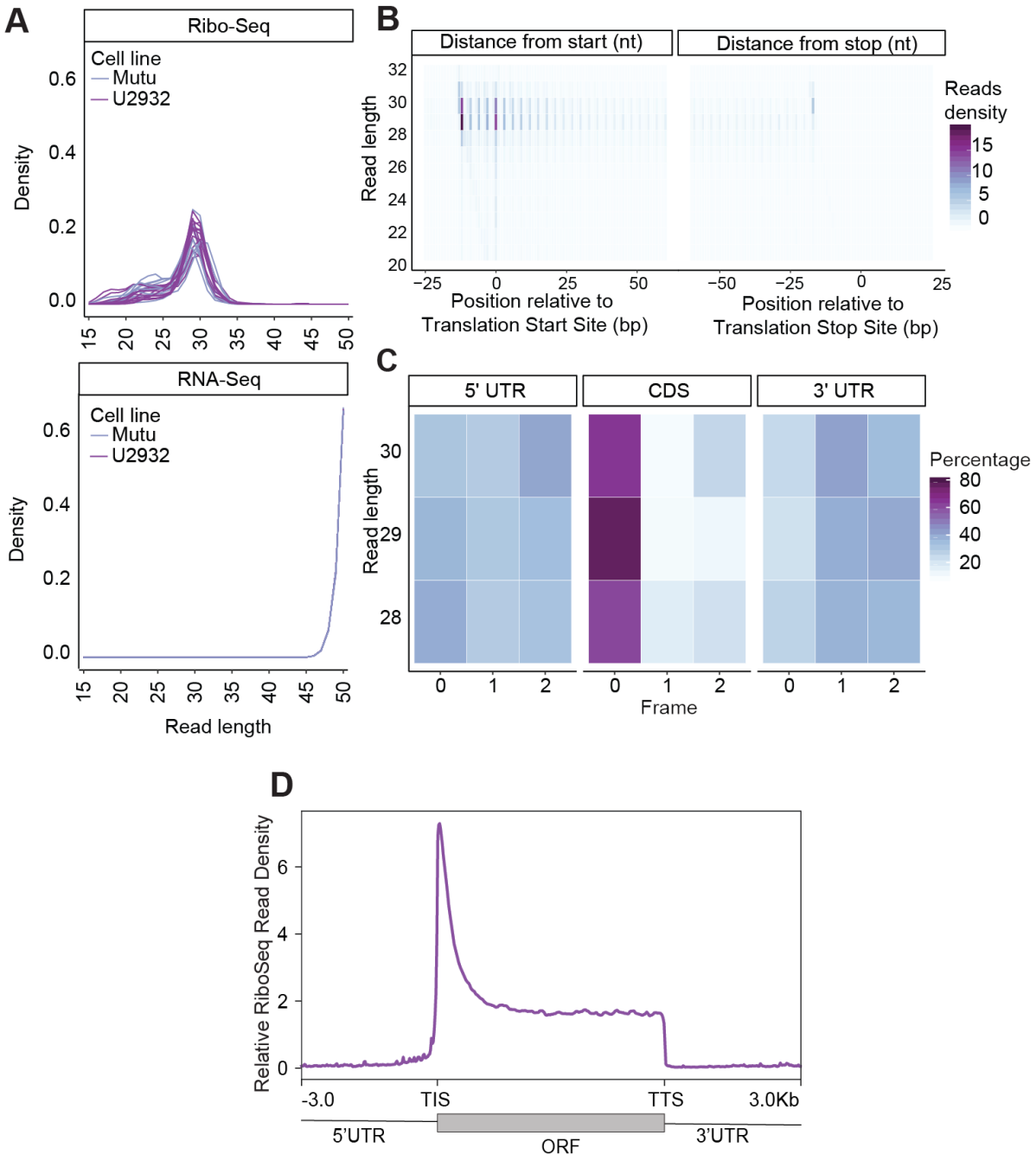

Figure 4.9: Quality control of Ribo-Seq dataset examining translational consequences of DDX3X depletion in U2932 and Mutu

A) Histogram showing the distribution of read length in Ribo-Seq and RNA-Seq experiments. Ribo-Seq shows a characteristic peak at 28-29 nt, which corresponds to the length of the ribosome protected mRNA fragment.

B) Heatmap of Ribo-Seq read frame usage by read length, showing read frame restriction in the 28-31nt ribosome protected fragments, characteristic of ribosome position on the mRNA transcript.

C) Heatmap of Ribo-Seq read frame usage by gene feature, showing how the characteristic read frame bias within the CDS but not the UTRs.

D) Metagene plot showing distribution of mapped Ribo-Seq reads to regions of the transcript, with the characteristic peak at the translation initiation site (TIS) and abrupt drop-off at translation termination site (TTS).

Translationally downregulated genes were enriched for components of the core translational machinery, in particular, protein constituents of the ribosome **(Figure 4.10 B-C)**. This was specific to the cytosolic ribosome as no significant difference in TE was observed for transcripts encoding mitochondrial ribosomal proteins (**Figure 4.11 B** p-value = $2.2 \cdot 10^{-16}$, Kolmorgorov-Smirnov test). The group of translationally downregulated genes also included ODC1, a known gene translationally regulated by DDX3X (Calviello et al., 2021). No significantly enriched terms were found in the translationally upregulated mRNA up or mRNA down groups. Genes in TE down group were more likely to be identified as iCLIP targets (**Figure 4.11 A-C**, p-value=$3 \cdot 10^{-14}$, Chi-squared test). Identified translationally regulated genes spanned broad range of mRNA expression levels.

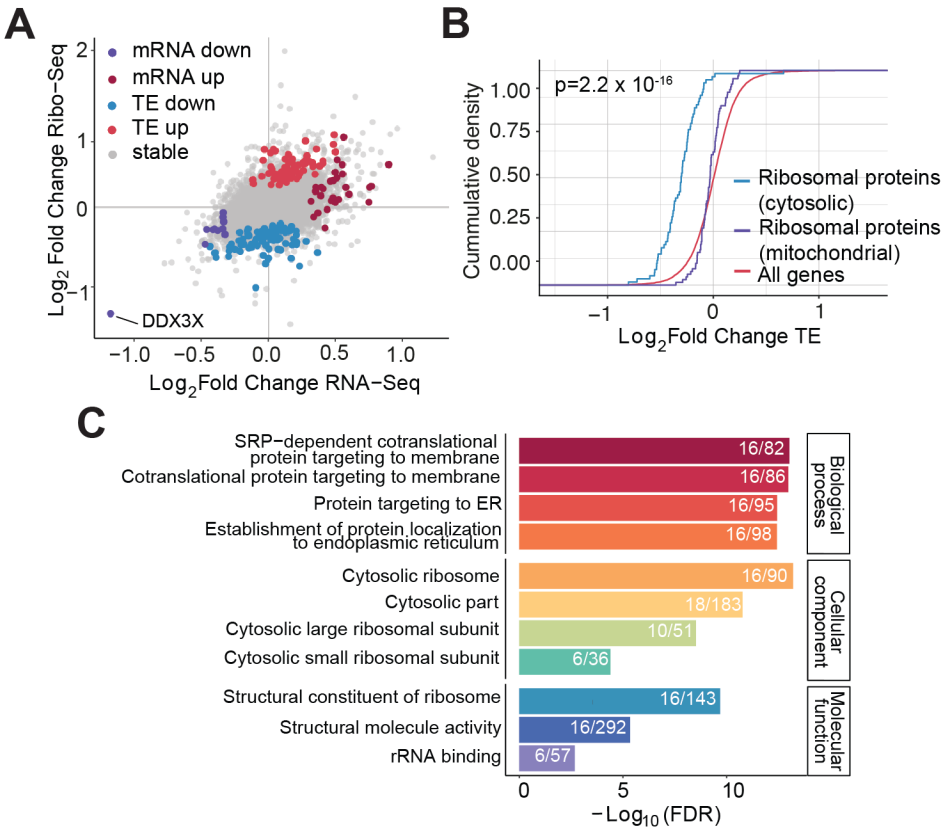

Figure 4.10: DDX3X-sensitive transcipts are enriched for the components of the cytosolic ribosome

A) Scatter plot comparing changes in mRNA abundance (RNA-Seq) with changes in ribosome footprint density (Ribo-Seq) following shRNA depletion of DDX3X in U2932. Data are from eight replicate knockdowns using two different shRNAs. Transcripts with altered translational efficiency (TE) or mRNA abundance are indicated by color.

B) Cumulative distribution of TE change in U2932 following shRNA depletion of DDX3X is plotted for genes encoding cytosolic or mitochondrial ribosome proteins, or all other genes. P-value calculated using Kolmogorov-Smirnov test.

C) Barplot showing Gene Ontology (GO) enrichment of genes with reduced TE following DDX3X depletion.

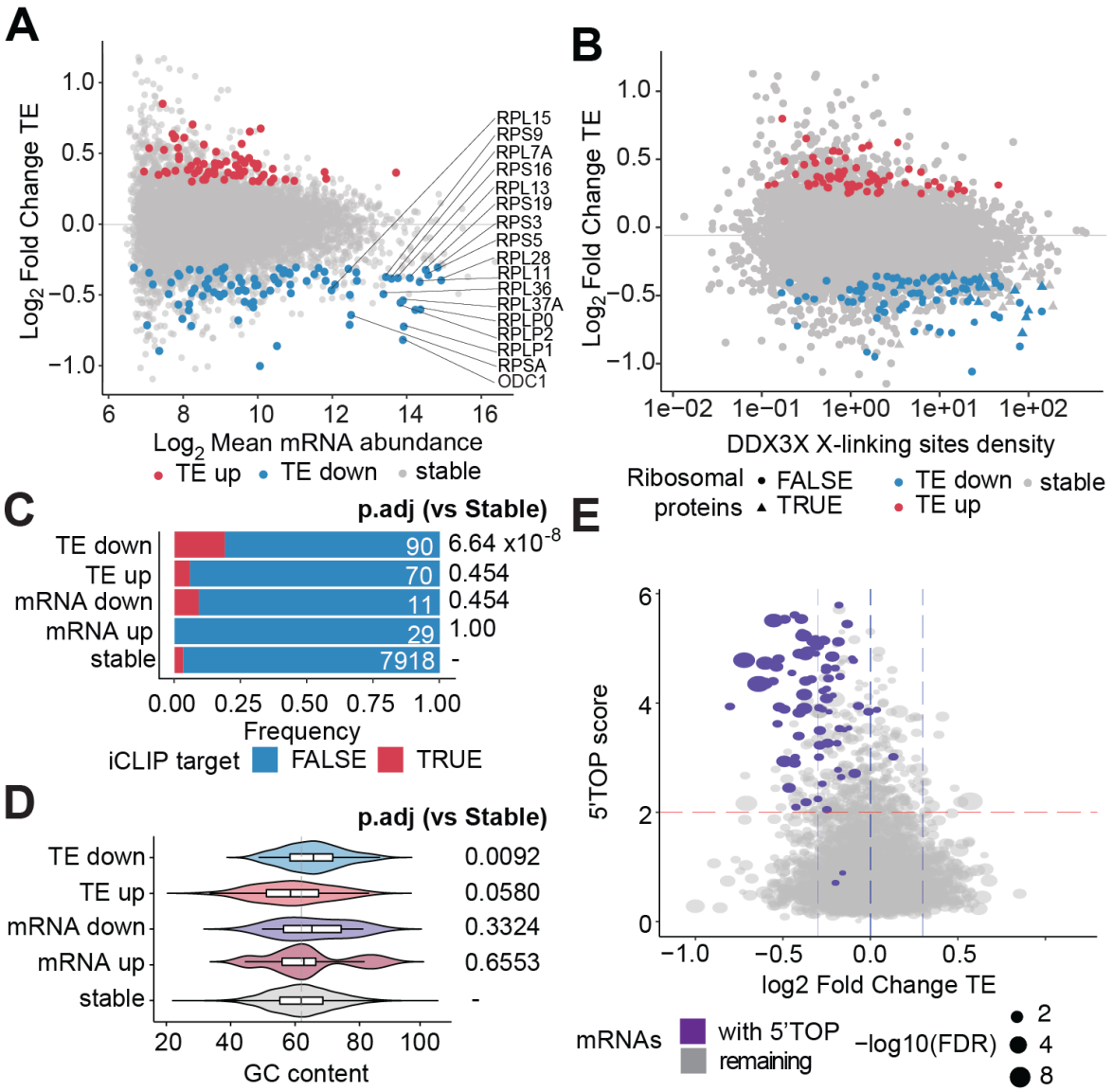

Figure 4.11: Differentially translated mRNAs following DDX3X depletion in U2932 are more likely to be bound by DDX3X

A) Scatter plot showing changes in TE plotted against mRNA abundance. Genes identified as differentially translated are coloured. Ribosomal proteins and ODC1 (a known DDX3X-regulated gene) are indicated.

B) Scatter plot showing changes in TE plotted against cross-lining density from iCLIP experiments.

C) Barchart showing the proportion of transcripts with differential translation identified as direct targets of DDX3X in iCLIP experiments. The number of genes within each category is indicated. Adjusted p-values (Fisher test) are shown and reflect the comparison of each category with stable genes.

D) Violin plot showing GC content distribution across different categories of differentially expressed genes. Adjusted p-values (Wilcoxon test) are shown and reflect the comparison of each category with stable genes.

E) Scatter plot showing the $\log_2$ fold changes in TE against 5′TOP score, which reflects the strength of the 5′TOP motif. Canonical 5′TOP genes are indicated.

The helicase activity of DDX3X may be linked to preferential translation of mRNAs with complex 5′UTR structure. I compared GC content and RNA folding energy of 5′UTR sequence in translationally regulated mRNAs to explore this hypothesis. Although translationally downregulated genes had slightly higher GC content (adjusted p-value = 0.0092, Wilcoxon rank sum test), the RNA folding energy was not significantly different between the regulatory groups (**Figure 4.11 D**). Because translational control of ribosomal proteins is linked to certain RNA motifs, such as 5 terminal oligopyrimidine (5′TOP) motif, I explored also this option. I downloaded the list of canonical 5′TOP and 5′TOP scores from a recent study surveying 5′TOP sequences in the human transcriptome (Philippe et al., 2020). A striking finding was that all canonical 5′TOP mRNAs, which include ribosomal proteins and few translation factors, showed a decreased translation efficiency following DDX3X depletion (p-value < 2.2e-16, Wilcoxon rank sum test), suggesting that DDX3X may regulate translation of its targets through 5′TOP motif (**Figure 4.11 E**).

Next, we validated the prediction of Ribo-Seq in three ways:

1. **Mass spectrometry of DDX3X shRNA U2932 cells:**
   The results of Ribo-Seq experiments were confirmed at the protein level with a tandem mass tag (TMT) mass spectrometry. In line with the Ribo-Seq results, the abundance of almost all cytosolic ribosomal proteins was reduced after shRNA depletion of DDX3X in U2932 cells **(Figure 4.12 A)**. GO analysis of proteins with reduced abundance in the DDX3X shRNA group showed enrichment for terms associated with protein synthesis, and the protein constituents of the ribosome, therefore reiterating the Ribo-Seq findings **(Figure 4.12 B-C)**. Interestingly, the depletion of R475S mutant DDX3X from Mutu had minimal effect on translation or mRNA abundance of specific transcripts, which supports the hypothesis of loss-of-function nature of this mutation**(Figure 4.12 B)**.

2. **ProteomeHD database of protein co-regulatory groups:**
   To examine the DDX3X interaction in a broader context, I queried Proteome HD database (Kustatscher et al., 2019). This repository integrates data from 5,288 mass-spectrometry runs spanning diverse human tissue types and biological conditions to create a map of the functional associations between co-expressed proteins. Examining co-regulation landscape of DDX3X showed 81 proteins, of these 27 were parts of the GO Cytosolic Ribosome group (adjusted p-value = $7.2 \cdot 10^{-42}$) and 8 were components of the core translation machinery **(Figure 4.13)**. Overall, GO terms with the strongest enrichment among DDX3X-interacting proteins were: RNP complex, mRNA metabolism, translation initiation and the cytosolic ribosome.

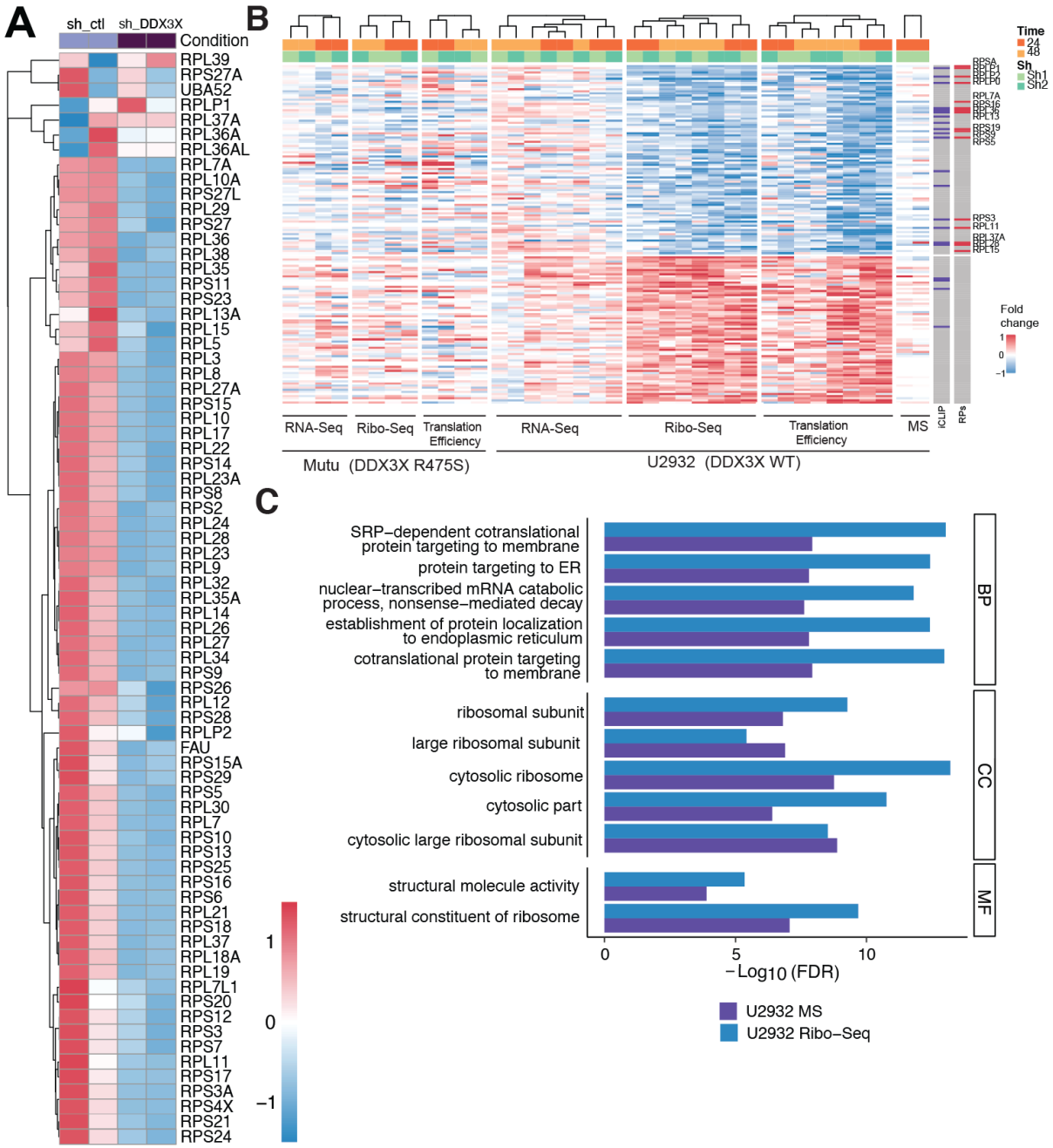

Figure 4.12: Validation of Ribo-Seq results with proteomic profiling of DDX3Xsh cells.

A) Heatmap showing altered abundance of ribosomal proteins in mass spectrometry analysis performed following shRNA DDX3X depletion in U2932.

B) Heatmap showing fold change in RNA-Seq, RiboSeq and TE across eight replicate knock-downs for all differentially translated genes. Protein abundance changes are shown (Mass Spec). DDX3X targets identified from iCLIP and genes encoding ribosomal proteins (RPs) are indicated by purple or red bands respectively.

C) GO terms enriched amongst proteins with reduced abundance in proteomic profiling (MS) following DDX3X depletion. Results from RiboSeq experiments are included for comparison.

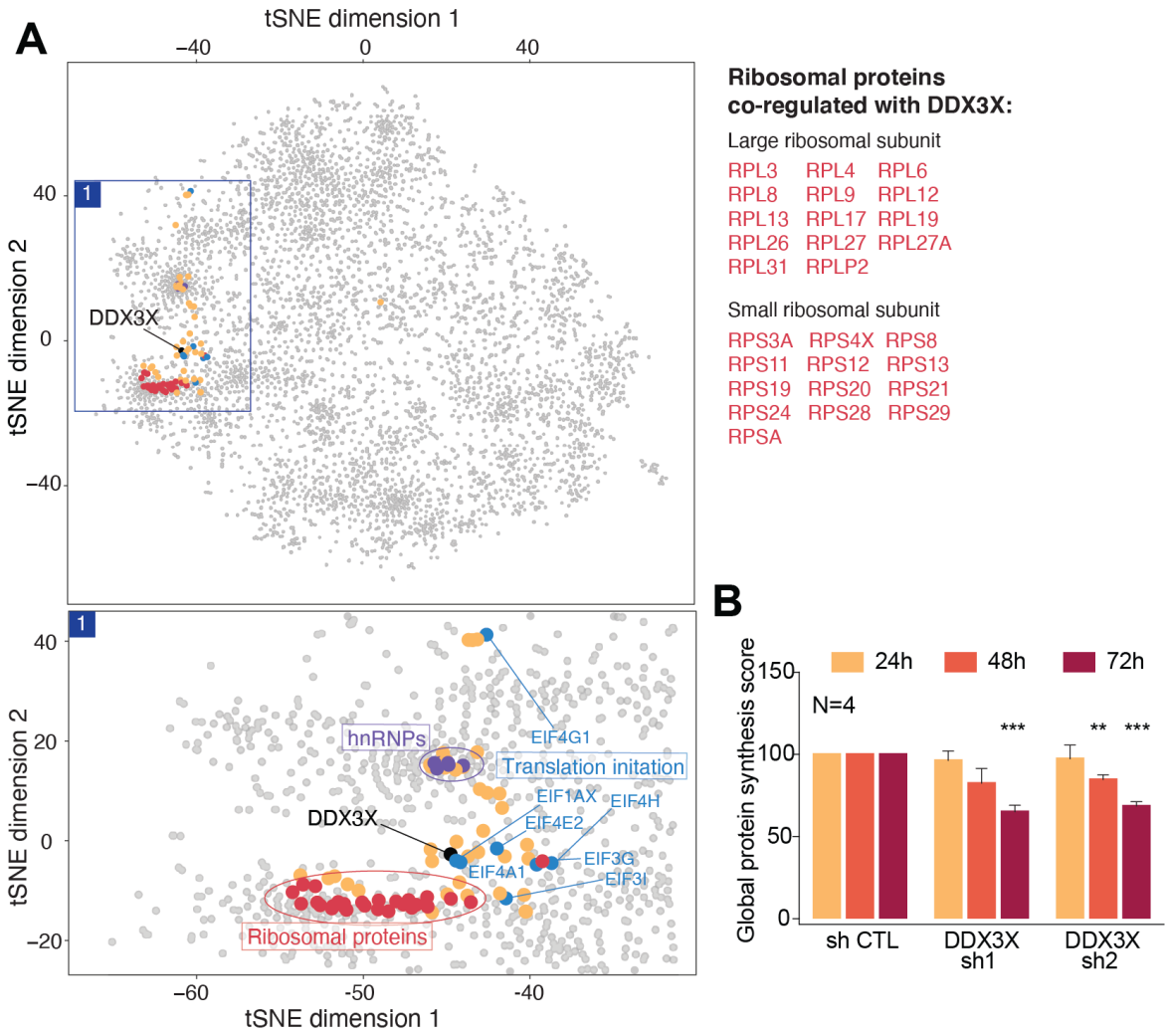

Figure 4.13: Validation of Ribo-Seq results with ProteomeHD analysis of proteins co-regulated with DDX3X and OPP-assay of DDX3X depleted cells

A) Map of the co-regulated human proteome plotted using data downloaded from Proteome-HD (Kustatscher et al., 2019). The 81 proteins identified as being statistically co-regulated with DDX3X are coloured. Ribosome proteins identified as being co-regulated with DDX3X are shown in red.

B) Barplot showing global protein synthesis quantified by OPP incorporation at the indicated time points following shRNA depletion of DDX3X in U2932 and normalized to control shRNA. Data shows mean+SEM, * $p < 0.05$, *** $p < 0.001$, ANOVA with multiple comparison testing, n=4 replicate experiments. Performed by Dr. Chun Gong.

3. **O-propargyl-puromycine (OPP) assay for measuring global translation rate:**
   Translational downregulation of the ribosomal proteins after shRNA depletion of DDX3X in U2932 was associated with a reduction in global synthesis rate **(Figure 4.13 B)**. A similar effect was observed in primary GC B-cells transduced with helicase-dead DDX3X mutants (Gong, Krupka et al., 2021).

Taken together, these results reveal that DDX3X promotes translation of transcripts encoding core components of translation machinery, in particular ribosomal proteins. The net effect of this is the regulation of global protein synthesis capacity.

4.2.3 Deregulation of MYC in primary GC B-cells increases ribosome biogenesis and triggers ER stress.

The observation that loss-of-function mutations in DDX3X affect ribosomal proteins synthesis, and hence decrease global translation load, contrasts sharply with the known role of MYC in promoting ribosome biogenesis. Relevant experimental system is key to understand molecular mechanism of lymphoma. GC B-cells are the cell-of-origin of BL, so it was essential to investigate the the transcriptional consequences of MYC overexpression in this system. Previous difficulties with genetic manipulations of *ex vivo* human B-cells means that this experiment has not been performed before. For this I analysed the RNA-Seq data I generated in chapter 3 (section 3.2) Transduction of MYC alone into primary GC B-cells triggers apoptosis, which is not the case for MYC transduction into established lymphoma cell lines. Therefore an experiment where MYC is overexpressed together with BCL2 (MYC-t2A-BCL2), should allow me to infer transcriptional response to MYC in the context of cell-of-origin of BL. Differential expression analysis comparing MYC-BCL2 with BCL2 alone revealed massive upregulation of a ribosome biogenesis signature. This was specific to MYC, as no such pattern was seen for BCL6-BCL2 cells. In addition, Gene Set Enrichment Analysis (GSEA) revealed strong upregulation of Unfolded Protein Response (UPR) signature. Moreover, ERN1, a key sensor of UPR, was one of the most significantly upregulated genes in MYC transduced GC B-cells.

UPR is an adaptive response to disturbance of the Endoplasmic Reticulum (ER) homeostasis, which results in the accumulation of misfolded proteins in the ER lumen. It triggers the expression of ER chaperones aiming to contain misfolded proteins through ER-associated degradation (ERAD) or commit the cell to apoptosis (Walter and Ron, 2011; Ruggiano et al., 2014; Hetz et al., 2015; Zhang et al., 2020). Molecular markers of UPR involve alternative splicing of transcription factor XBP1 initiated by ERN1 and phosphorylation of eIF2α by EIF2AK3 (PERK) (Ron and Walter, 2007).

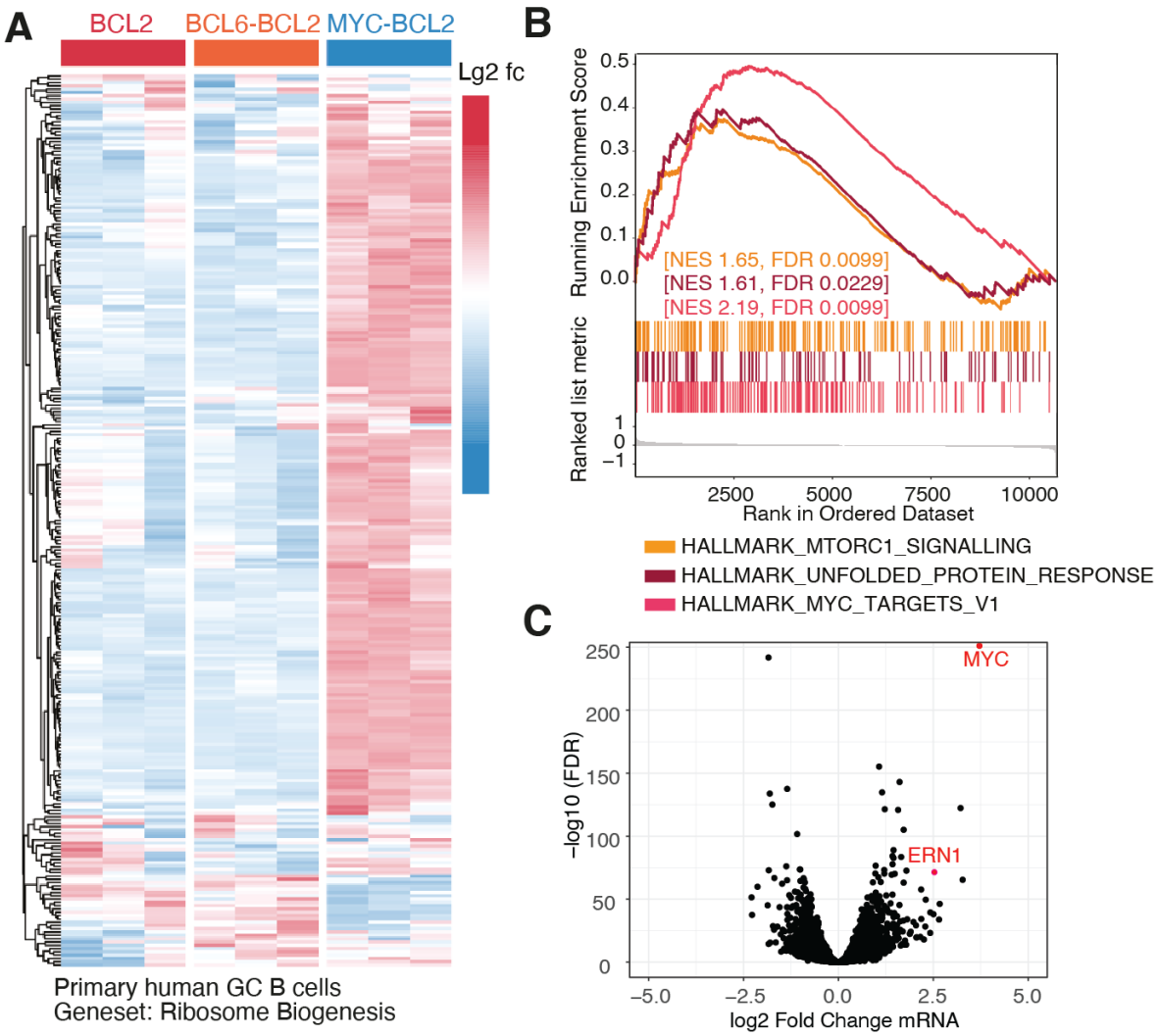

Figure 4.14: Deregulation of MYC in primary GC B-cells increases ribosome biogenesis and triggers Unfolded Protein Response

A) Heatmap showing mRNA expression level of genes belonging to the gene set Ribosome Biogenesis (GO: 0042254) in human GC B-cells transduced with BCL2, BCL6-2A-BCL2, or MYC-2A-BCL2

B) Gene set enrichment analysis (GSEA) of RNA-seq from human GC B-cells transduced with MYC-2A-BCL2 compared to BCL2 alone, showing enrichment of gene sets related to MYC, UPR, and mammalian target of rapamycin complex 1 (mTORC1) signalling. Genes ordered according to DESeq2 test statistics (decreasing order).

C) Volcano plot ($\log_2$ fold change against -$\log_{10}$(FDR) from differential expression analysis MYC-2A-BCL2 compared to BCL2 alone). The position of MYC and ERN1 indicated.

Elevated levels of phosphorylated eIF2α and spliced isoform of XBP1 in primary GC B-cells transduced with MYC-t2A-BCL2 were confirmed experimentally by Dr. Chun Gong. UPR turned out to be, at least partly, related to MYC-induced apoptosis. Primary GC B-cells transduced with MYC alone and treated with rapamycin, an allosteric inhibitor of mTORC1, had significantly lower protein synthesis associated with a modest reduction in apoptosis rate. This experiment was performed by Dr Chun Gong (Gong, Krupka et al., 2021).

These results show that deregulation of MYC in primary GC B-cells is associated with a higher protein synthesis rate and induction of ER stress response. Loss of DDX3X, by limiting global protein synthesis might, protect cells from proteotoxic stress, thus alleviating MYC-induced apoptosis. Support for this hypothesis comes from a series of experiments described in the next section.

4.2.4 DDX3X mutation interferes with endoplasmic reticulum stress response

4.2.4.1 DDX3X R475C mutation is associated with suppression of unfolded protein response in U2932 cells

A combination of computational and experimental evidence allowed us to hypothesise that loss-of-function mutations in DDX3X may play a similar role to rapamycin treatment in protecting cells from MYC-induced apoptosis. If deregulation of MYC in primary GC B-cells is associated with proteotoxic stress and apoptosis, DDX3X mutations may lower translation load, alleviate ER stress, thus decreasing apoptosis rate.

To interrogate the regulatory network of DDX3X mutation in lymphoid cells, I analysed RNA-Seq data from an experiment performed by Dr. Gong comparing CRISPR–edited clones of U2932 cells expressing R475C helicase mutant DDX3X. Control clones with a synonymous mutation in endogenous DDX3X were created in parallel. Differential expression analysis comparing CRISPR-Cas9 DDX3X clones revealed strong downregulation of the key regulators of ER stress response - ERN1 and XBP1. Moreover, the mRNA expression pattern of R475C-edited clones resembled the profile of samples with the strongest depletion of DDX3X following shRNA treatment. GSEA analysis of DDX3X-mutant clones and RNA-Seq of shDDX3X depleted cells also revealed striking overlap. In both comparisons 'MYC targets V1', 'MTORC1 signalling' and 'Unfolded Protein Response' terms were among the most downregulated gene sets.

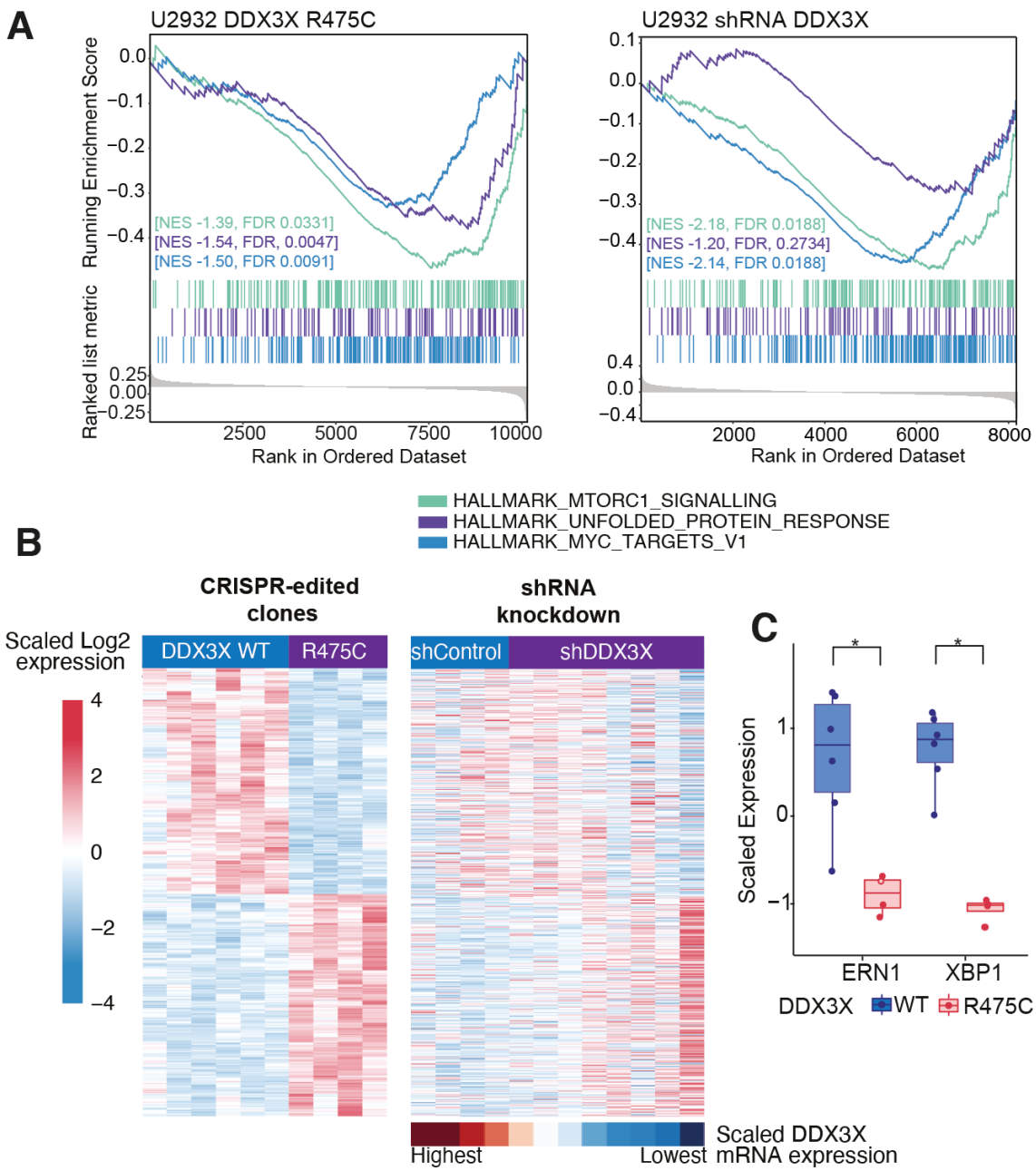

Figure 4.15: DDX3X R475C mutation is associated with downregulation of genes associated with unfolded protein response in U2932 cells

A) GSEA analysis of RNA-seq data comparing DDX3X R475C-edited or control clones (left), and DDX3X shRNA knockdown experiments (right). Genes ordered according to DESeq2 test statistics (decreasing order).

B) Heatmap showing genes that are differentially expressed between control and homozygous R475C edited clones (left). The same genes are shown for shRNA knockdown (right). For shRNA experiments samples are ordered left to right by DDX3X mRNA expression. The top bar indicates the expression of DDX3X mRNA showing how stronger knockdown recapitulates the signature seen in R475C-edited clones.

C) Boxplot showing relative expression of the Unfolded Protein Response (UPR) marker transcripts ERN1 (encoding IRE1) and XBP1 mRNA in RNA-seq from DDX3X R475C-edited clones. Statistical significance from differential expression analysis (DESeq2).

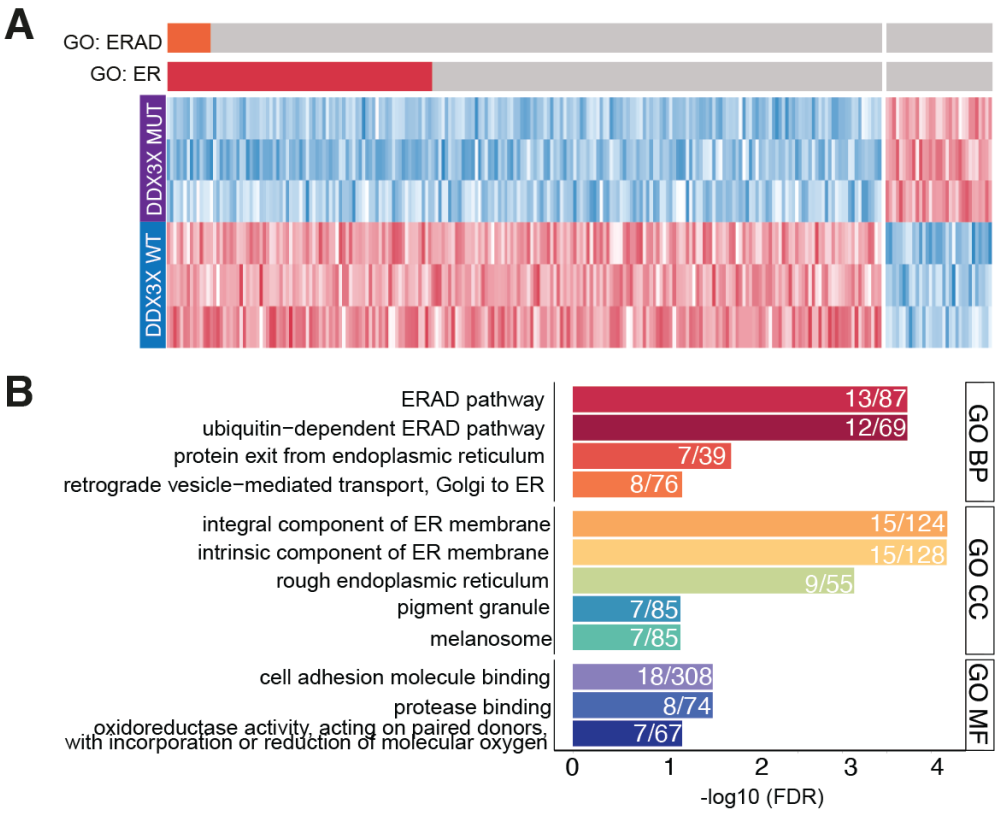

Figure 4.16: DDX3X R475C mutation is associated with downregulation of proteins associated with ER stress in U2932 cells

A) Heatmap showing proteins with altered abundance in proteomic profiling of DDX3X R475C-edited clones. Proteins included in the Gene Ontology (GO) terms Endoplasmic Reticulum (ER), GO: 0005783, and ER-associated protein degradation pathway (ERAD), GO: 0036503, are indicated by red and orange highlighting, respectively.

B) Barplot showing the statistical significance of top GO terms enriched among proteins with decreased expression in DDX3X R475C-edited clones. BP, biological process; CC, cellular component; MF, molecular function.

Given the evidence of DDX3X driving gene expression changes predominantly at the level of translation, the R475C mutant and control clones were subjected to proteomic profiling. In line with previous experiments, GO terms linked to protein processing in the ER and ER stress were enriched in protein downregulated in R475C mutant clones. Nearly one-third of downregulated proteins were related to GO terms associated with the ER or ER stress.

I conclude that the R475C mutation can be considered loss-of-function and is associated with suppression of expression of genes related to unfolded protein response.

4.2.4.2 DDX3X mutation is associated with suppression of unfolded protein response in BL patients

In an orthogonal approach, I examined gene expression profiles in human biopsy samples obtained from two published BL datasets (Grande et al., 2019; Schmitz et al., 2012). I downloaded raw FASTQ files from Sequence Read Archive, performed quality check, alignment to the reference genome and differential expression and GSEA analysis comparing DDX3X mutant to wild-type samples. Expression pattern of DDX3X mutant cases mirrored the changes observed in our CRISPR-edited clones: reduced expression of 'MYC targets', 'MTORC1 signalling' and 'Unfolded Protein Response' sets in GSEA accompanied by strong downregulation of ERN1 and XBP1 mRNA levels.

The link between DDX3X mutation and ER stress response was confirmed experimentally by Dr. Chun Gong. Co-transduction of primary GC B-cells with MYC and DDX3X mutant recapitulated the pattern observed after rapamycin treatment: loss-of-function DDX3X mutations (R475C and K230E) abrogated both the MYC-induced apoptosis as well as the increase in the global translation rate caused by MYC. DDX3X mutation was also able to alleviate ER stress response induced by treatment with thapsigargin (Gong, Krupka et al., 2021), which inhibits the sarco/endoplasmic reticulum Ca2+ ATPase (SERCA) thereby inducing ER stress .

Taken together, I conclude that loss-of-function DDX3X mutations can counteract the effects of MYC to drive global protein synthesis and trigger proteotoxic stress. This also reveals a potential vulnerability of MYC-driven lymphomas to drugs inducing ER stress.

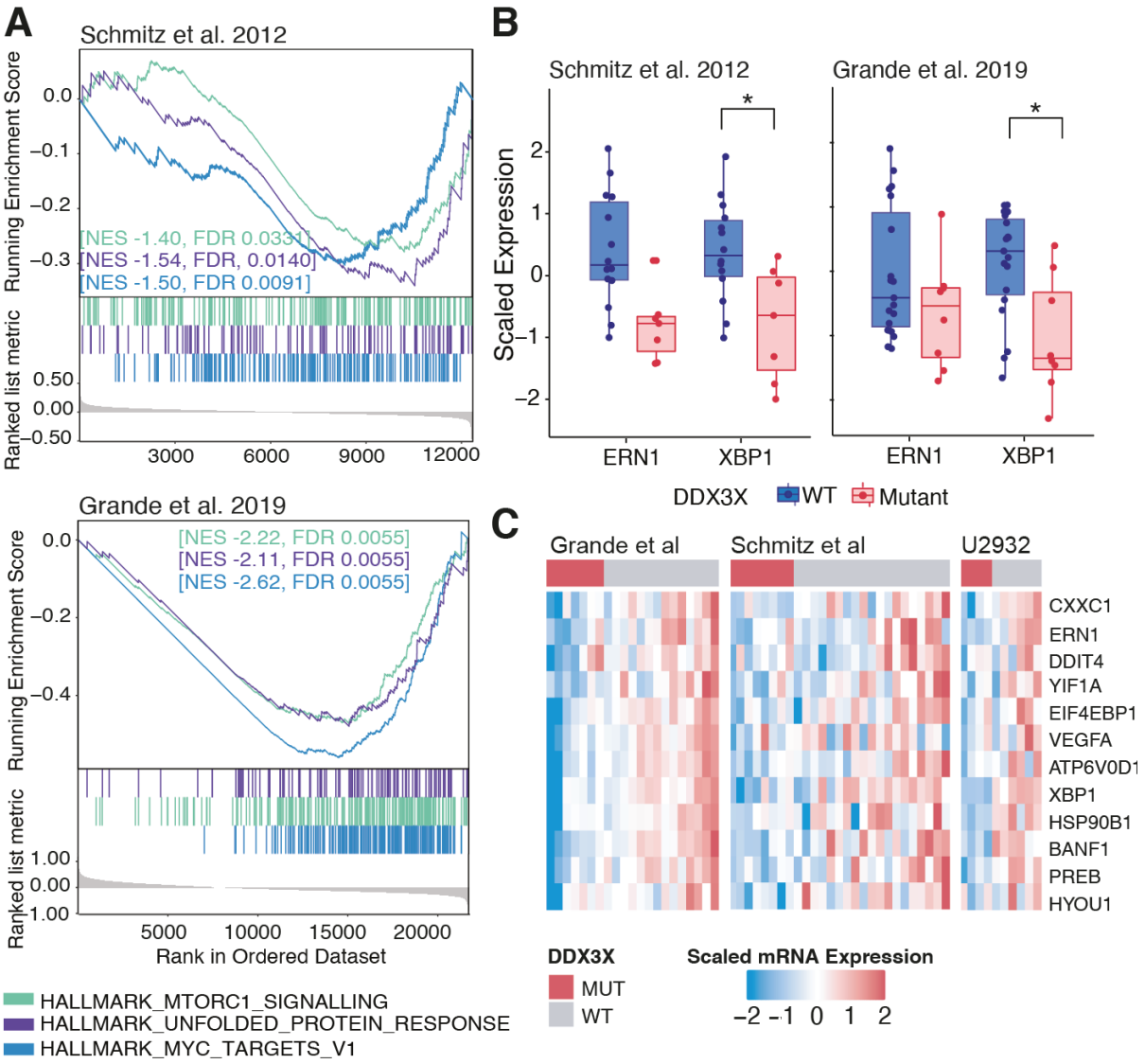

Figure 4.17: Downregulation of genes associated with unfolded protein response in BL patients with DDX3X mutation

A) GSEA analysis of RNA-seq data from the indicated studies reanalyzed to compare cases of sporadic BL with either WT or mutant DDX3X. Genes ordered according to DESeq2 test statistics (decreasing order).

B) Relevant gene sets downregulated in the presence of mutant DDX3X or the relative abundance of the Unfolded Protein response (UPR) transcripts ERN1 and XBP1. Statistical significance is from DESeq2.

C) Heatmap showing mRNA expression in RNA-Seq of GSEA core enrichment genes in the gene set Hallmark Unfolded Protein Response in BL biopsies in two published Burkitt lymphoma RNA-Seq data sets and in U2932 DDX3X R475C edited clones.

4.2.5 Up-regulation of DDX3Y in established tumours rescues loss of DDX3 helicase activity

Context-dependent effect of DDX3X activity was a recurrent finding in this and previous studies. While loss-of-function DDX3X mutation was beneficial for primary GC B-cells co-transduced with MYC, it triggered apoptosis in established lymphoid cell lines (Gong, Krupka et al., 2021). Because upregulated translation rate is a feature of established tumours, and the ability to increase global protein synthesis is essential for MYC-driven lymphomas (Barna et al., 2008), we hypothesised that there must be a way to compensate for the reduced translation capacity during the latter stages of lymphomagenesis.

Both BL and MHG subtypes of DLBCL have strongly skewed sex ratios in favour of males and DDX3X mutations are more frequent in male-skewed cancers (Alkallas et al., 2020; Dunford et al., 2017). DDX3X shares almost 92% amino–acid and 70% nucleotide sequence identity with the Y–chromosome homologue, DDX3Y. At the functional level, it is redundant with DDX3X in regulating protein synthesis (Venkataramanan et al., 2020). Firstly, DDX3**Y** can rescue a decrease in total translation after depleting DDX3**X** and secondly, the translation profile of single transcript in male-derived colorectal cancer HCT 116 cells is indistinguishable (Venkataramanan et al., 2020). DDX3Y is widely transcribed in many adult tissue types, but it is not expressed at the protein level anywhere except spermatogonia (Ditton et al., 2004; Foresta et al., 2000a; Rauschendorf et al., 2011). To establish the status of DDX3Y expression in lymphoid cells, I examined changes in mRNA abundance in a previously published RNA-Seq dataset including different states of GC B-cells (processed by me from FASTQ files) (Caeser et al., 2019). I did not observe significant changes in DDX3X mRNA comparing freshly isolated GC B-cells, established lymphoma cell lines, as well as transduced and transformed GC B-cells. Finally, no significant difference in DDX3X mRNA level was observed between DDX3X mutant and wild–type samples in male BL patients (Grande et al., 2019). However, the coverage of ribosomal footprints in the DDX3Y region from Mutu cells was similar to known protein coding regions with similar expression level. In line with that, immunoblotting for DDX3Y in male lymphoma cell lines performed by Dr. Chun Gong revealed strong protein expression. The same was observed in patient-derived BL xenografts and 5 primary biopsies from male patients (Gong, Krupka et al., 2021). This is consistent with previous works proposing that DDX3Y protein expression in testis is regulated predominantly at the level of translation through alternative usage of translation initiation site in 5′UTR of DDX3Y (Jaroszynski et al., 2011). I questioned whether DDX3Y translation might be directly influenced by DDX3X. However, I found no evidence from the iCLIP or Ribo-Seq experiments that the DDX3Y transcript was a direct target of DDX3X. Nevertheless, the link between loss-of-function mutation in DDX3X and protein expression of DDX3Y

remains strong. *In vivo* tumorigenesis experiment performed by Dr. Chun Gong to validate this observation showed that primary GC B-cells with deleted DDX3X, transduced with MYC-2A-BCL2 and implanted subcutaneously in Matrigel into immunodeficient mice, forms tumours and start to express DDX3Y protein seven weeks after injection. This suggests that DDX3Y expression at the protein level is unique to transformed B-cells. Although indirectly, the expression of DDX3Y is linked to loss of DDX3X, but the exact mechanisms of this is unclear.

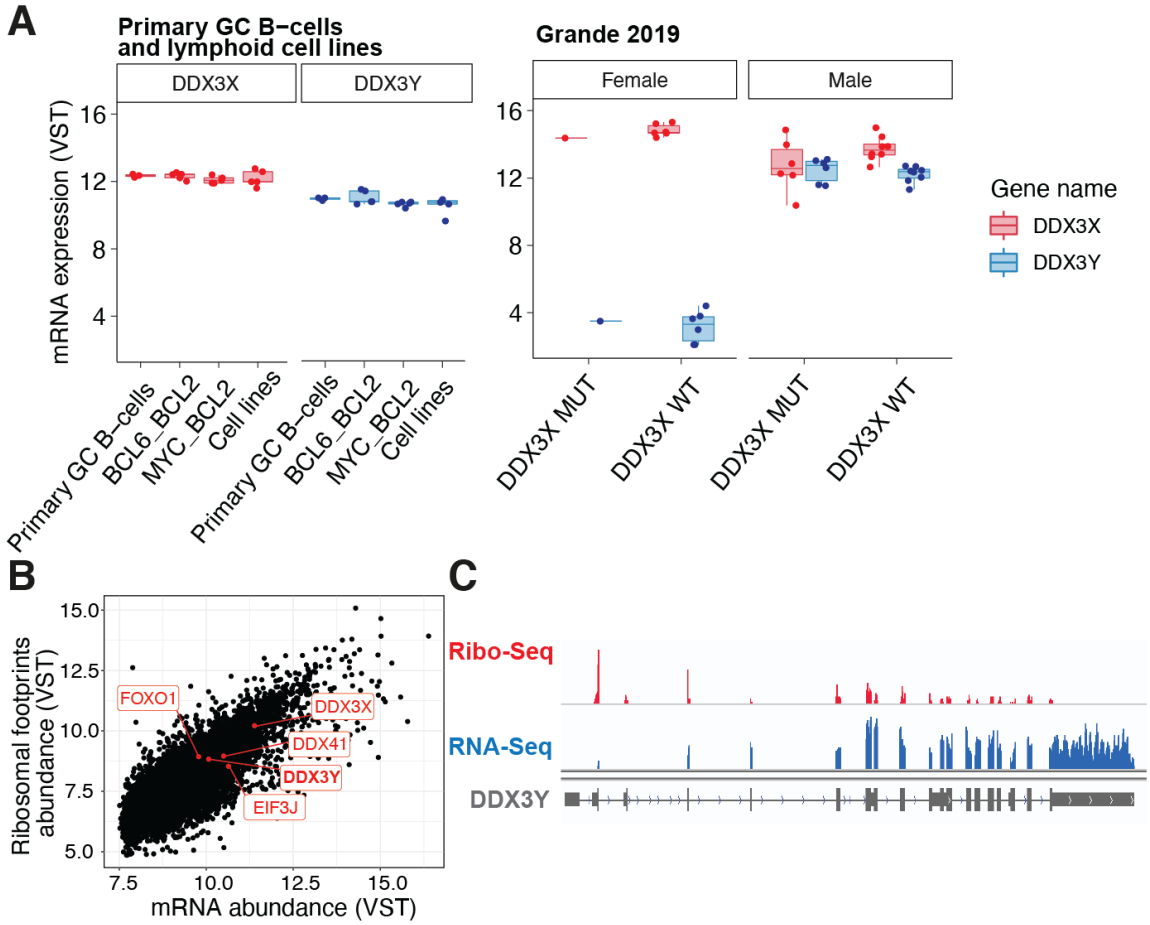

Figure 4.18: Expression of DDX3Y and DDX3X in primary GC B-cells and lymphoma cell lines

A) Boxplot showing the mRNA expression level of DDX3X and DDX3Y in Primary GC B-cells, GC B-cells transduced with BCL6-BCL2 or MYC-BCL2 constructs and in lymphoid cell lines (left panel) and in BL patients for males and females separately (right panel). VST, Variance Stabilising Transformation

B) Scatter plot of mean mRNA and ribosomal footprints abundance (VST transformed) showing the expression level of DDX3Y in comparison to other genes. DDX3X, DDX41 (ubiquitously expressed RNA helicase), FOXO1 (BL oncogene) and EIF3J (translation initation factor) were indicated for comparison.

C) Coverage plot of Ribo-Seq and RNA-Seq reads in the DDX3Y region in representative samples from Mutu cells

4.3 Discussion

Main findings

Although recurrent DDX3X mutations have been identified in a variety of malignancies, the molecular role of DDX3X in malignancy remains puzzling. The versatile function of DDX3X in regulating multiple stages of RNA biogenesis is mirrored by its complex and context-specific role in tumorigenesis - both oncogenic and tumour suppressor activity has been reported. This study discusses the function of DDX3X in MYC-driven B-cell lymphoma revealing functional cooperation between MYC and mutant DDX3X. Deregulation of MYC expression, through translocation to highly expressed immunoglobulin loci or mutations increasing protein stability, is central to the development of Burkitt Lymphoma and Double Hit or Molecular High-Grade subtype of DLBCL. All three diseases share the germinal centre origin and are associated with poor clinical prognosis. The model proposed here highlight the vulnerability of MYC-driven lymphoma to proteotoxic stress opening an attractive opportunity for therapeutic intervention. By integrating data from different high-throughput sequencing techniques I show that the effect of DDX3X mutations is mainly mediated by changes in translation of selected transcripts, predominantly ribosomal proteins, which in turn controls global translation intensity. The requirements for translation load change during lymphomagenesis, so that the state of decreased translation capacity might not meet the demands of a fully established tumour. By sequential deregulation of DDX3X and DDX3Y, it is possible to buffer global protein synthesis to support these changing needs. Therefore drugs that disrupt this delicate balance of translation load and proteotoxic stress may prove effective against MYC-driven lymphoma.

Context specific role of DDX3X in malignant cells

DDX3X mutations have been widely studied in medulloblastoma, the most common brain tumour in children, where they are associated with Wingless (WNT) and Sonic hedghog (SHH) subtypes (Northcott et al., 2017; Patmore et al., 2020). The pattern of DDX3X mutations in medulloblastoma differs from what I have observed in BL, which suggests a different molecular role of DDX3X mutant in the two tumours. While nonsense and frameshifts mutations are frequent in BL, they are never observed in medulloblastoma. Moreover, amplification of MYC family genes, which is a hallmark of Group 3 and Group 4 medulloblastoma subtype is never accompanied by mutations in DDX3X (Northcott et al., 2017), indicating that the cooperative effect between DDX3X and MYC is also context-specific. DDX3X has been shown to regulate brain development and mediates tumour suppressing stress and inflammasome response (Patmore et al., 2020). Medulloblastoma-

associated DDX3X mutations promote oncogenic WNT signalling directly in helicase-independent fashion through an association with CSNK1E protein (Cruciat et al., 2013; Pugh et al., 2012) and indirectly, through prevention of pyroptosis, inflammatory apoptosis mediated by inflammasome activation coupled with WNT signalling (Huang et al., 2020; Samir et al., 2019; Patmore et al., 2020). I did not find compelling evidence of the association between DDX3X and CSNK1E in B-cells. Other previously reported individual targets of DDX3X include ODC1 (Van Steeg et al., 1991; Calviello et al., 2021), KLF4 (Cannizzaro et al., 2018) and MITF (Phung et al., 2019). Of which, only ODC1 was found to be DDX3X sensitive in lymphoid cells highlighting the strong tissue-specific context of DDX3X activity. In contrast, I collected multiple orthogonal lines of evidence to support the regulation of ER stress and protein synthesis as the dominant phenotype of DDX3X mutations in MYC-driven lymphomas. This, in light of these data, is achieved through the regulation of translation of selected transcripts encoding the core components of the translation machinery. DDX3X–DDX3Y axis maintains the balance between translation load and proteotoxicity that is stage-specific and allows the germinal centre B-cells for adaptation to deregulated MYC expression. This study, however, does not allow me to exclude the possibility that loss-of-function mutations in DDX3X exert additional functions that may converge with functions of DDX3X reported in other tissue types.

Role of DDX3X as auxiliary translation factor

DDX3X is reported to play multiple roles in RNA biology, however, this study indicates that in primary GC B-cells and lymphoma cell lines, DDX3X acts predominantly at the level of translation. The role of DDX3X in regulating protein synthesis has been studied previously revealing both activating and repressing functions. Although this study does not define the precise mechanism of translational repression upon loss of DDX3X in B-cell lymphoma, at least two models could potentially explain this phenomenon.

Firstly, DDX3X is known to act as an auxillary translation factor directly interacting, independently on RNA, with eIF3, and 40S subunit of the ribosome (Geissler et al., 2012; Lee et al., 2008b). DDX3X induces conformational changes of the 43S pre-initiation complex promoting the release of translation initiation factors and joining the 60S ribosomal subunit leading to 80S complex formation (Geissler et al., 2012). In line with this study, the role of DDX3X in translation seems to be limited mainly to the initiation stage. In lymphoid cells, we observed that DDX3X protein co-immunoprecipitates with almost all components of the eIF3 complex and DDX3X iCLIP binding profile resembles the pattern of known translation initiation factors (Calviello et al., 2021). In fact, DDX3X has not been found in polysome fractions of human hepatoma cells by Geissler et al. (2012) reinforcing translation initiation as the main point of DDX3X mediated translational control.

Although DDX3X depletion affects translation intensity globally, it is worth mentioning

that Geissler et al. (2012) found that only about 50 % of the newly assembled ribosomes contained DDX3X, which suggests that regulation of translation by DDX3X might be limited only to a subset of transcripts. In fact, as an RNA-helicase, DDX3X can participate in the unwinding of highly structured regions within 5′UTRs, thus, facilitating translation initiation, but the evidence underlying this activity is puzzling. Soto-Rifo et al. (2012) and Calviello et al. (2021) work on DDX3X showed that transcripts with complex 5′UTR, originating from viral or host transcription respectively, are sensitive to DDX3X depletion. This is in contrast to Geissler et al. (2012), who argue that depletion of DDX3X had a substantial impact on the translation of all studied viral transcripts regardless the complexity of the leader sequence.

Almost complete loss of helicase activity for BL-associated DDX3X mutant suggests that the molecular function of DDX3X in BL is helicase-dependent. This would argue in favour of the 5′UTR structure as a key determinant for DDX3X-sensitive transcript. Although, DDX3X-sensitive transcripts had slightly higher GC content, their RNA folding energy of their 5′UTR was not significantly different from other genes suggesting that other factors may be involved.

An alternative hypothesis linking DDX3X with translation regulation of the core components of translation machinery is 5′TOP-mediated regulation. 5′ terminal oligopyrimidine (5′TOP) motifs are situated in the 5′UTR of selected transcripts. The translation of 5′TOP mRNAs is inhibited by LARP1, an RNA-binding protein, which is inactivated by mTORC1 complex. There is a possibility that DDX3X competes with LARP1 to control translation of 5′TOP transcripts or acts as 5′TOP translation modulator. Apparently, almost all canonical 5′TOP mRNAs have negative TE changes after DDX3X knock-down in U2932, but further experiments are needed to confirm if translational control by DDX3X is mediated through 5′TOP and/or mTOR control.

Different DDX3X expression levels may promote different cellular responses

The observation of *dose-dependent* mRNA expression profile of DDX3Xsh cells, which resembles the profile of R475C cells only when depletion is the strongest could suggest that even small amount of DDX3X is sufficient to maintain translation of target transcripts. Dual, dose-dependent role of DDX3X in translation has been reported previously. While moderate overexpression of DDX3 stimulated translation Geissler et al. (2012), massive overexpression lead to a halt in global translation. Given the versatile function of DDX3X in regulating different aspects of RNA biology, there is a possibility that different functions require precise titration of DDX3X levels in the cell.

In addition to the regulation of translation initiation, studies of the yeast homologue of DDX3X, Ded1, show that it is able to nucleate stress granules (SGs) formation (Hilliker et al., 2011). While sequestration of selected transcripts in SGs does not require Ded1

helicase activity, the reverse process of releasing mRNA into the cytoplasm does so, thereby promoting their translation (Hilliker et al., 2011). Indeed, SGs formation in human cells was reported in studies using overexpressed mutant DDX3X (Lennox et al., 2020a; Valentin-Vega et al., 2016). Valentin-Vega et al. (2016) argue that low complexity domains flanking the helicase core of DDX3X are the main effector of SGs assembly. This could explain the observed toxicity of supraphysiological levels of either wild-type or mutant DDX3X (Lennox et al., 2020a; Valentin-Vega et al., 2016).

Pro-oncogenic properties of decreased translation load

The process of oncogenesis is typically associated with an increased translation load that meet the demands of deregulated cell cycle and growth (Ruggero, 2013). However, in several malignancies, an opposite strategy has evolved. In neuroblastoma, the most common extracranial tumour in children, adaptation to limited nutrient supply is mediated by the activation of eEF2-kinase, which suppress translation elongation (Leprivier et al., 2013). It is also interesting that this effect was found predominantly in MYC-N-driven tumour model (Delaidelli et al., 2017) suggesting that tight regulation of translation load might be a general feature of malignancies deregulation of the MYC gene family. RUNX1 deficiency, which is common in myelodysplastic syndrome and acute myelogenous leukaemia (AML), reduces the rate of ribosome biogenesis making the haematopoetic stem cells (HSCs) resistant to stress, thus, providing a competitive advantage over normal HSCs (Cai et al., 2015). Another example comes from a mouse model of medulloblastoma, where an activating mutations in PERK are essential for premalignant cells to decrease translation load, which is restored at later stages of tumour formation (Ho et al., 2016). PERK (EIF2AK3) is a key activator of the integrated stress response, which target 5′cap translation initiation. Similarly to our model, lowering of global translation at early stages of tumorigenesis decreases ER stress and protects cells from apoptosis (Ho et al., 2016).

The conclusion that suppressed translation is advantageous for MYC-induced lymphomagenesis might be inconsistent with studies using existing mouse models of MYC-driven lymphoma. Haploinsufficiency for *Rpl24* in *Eμ-Myc* mouse, which overexpress *Myc* in the B-cell compartment, decreases global protein synthesis and delayed the lymphoma onset (Barna et al., 2008). This is radically different to the two-stage model proposed in this study. Genetic deletion of *Rpl24* induces permanent suppression of translation, whilst loss of DDX3X can be rescued by expression of DDX3Y protein. Thus the level of translation can be tuned to changing requirements of the newly forming tumour. This hypothesis relies strongly on the assumption that DDX3X and DDX3Y are redundant in their molecular function. Germline heterozygous mutations in DDX3X has long been linked to cortical malformations, intellectual disability and autism-spectrum disorders in females (Lennox et al., 2020b). Males with DDX3Y mutation, in turn, present with

infertility, but no cognitive dysfunction has been observed suggesting that DDX3X cannot compensate for loss of DDX3Y (Ferlin et al., 2003; Foresta et al., 2000b). This might be attributed to the difference in expression pattern between tissue types, however, I cannot exclude the possibility that the ectopic expression of DDX3Y in transformed B-cells is more than just rescuing DDX3X insufficiency. Nevertheless, the evidence supporting the overlap in molecular role of DDX3X and DDX3Y is rich. Firstly, DDX3X and DDX3Y are 92% identical in the amino-acid sequence. Secondly, recently published work reveal that substitution of DDX3X with DDX3Y completely recapitulates gene expression profile at the level of transcription and translation (Venkataramanan et al., 2020). Finally, we observed that DDX3Y bound the same mRNA targets as DDX3X and had a similar influence on regulating global protein synthesis (Gong, Krupka et al., 2021). The ability to reactivate DDX3Y protein expression may explain the observed sex bias towards males in BL and other cancers with recurrent DDX3X mutations. Given the importance of restored protein synthesis capacity in mature tumours, a natural question arises: what compensatory mechanism exists in female lymphomas, who also (but at much lower frequencies) present with DDX3X mutation. A review of mutation co-occurrence in BL exome sequencing data (Grande et al., 2019) shows that mutations in SIN3A are enriched (24 % vs. 6%; p-value < 0.05, Fisher Exact test) in the presence of DDX3X mutation. SIN3A is a transcriptional corepressor that antagonises MYC directly interacting with MXD1-MAX heterodimers (Nascimento et al., 2011). This observation allows me to speculate that SIN3A inactivation at the latter stages of lymphomagenesis could increase MYC activity and thereby global translation capacity. Other complementary mechanisms, genetic or non-genetic, restoring protein synthesis or increasing MYC activity are also possible. In contrast, this problem does not exist in males. Ectopic expression of DDX3Y has been found in all investigated cell lines. Moreover there are almost no mutations in DDX3Y across published lymphoma datasets, which suggests that functional DDX3Y is important for successful lymphomagenesis.

The importance of DDX3X/DDX3Y axis in MYC-driven lymphoma opens new therapeutic opportunities

The results presented here reinforce therapeutic relevance of drugs targeting ER stress for the treatment of MYC-driven lymphoma. Firstly, they support the vulnerability of the lymphomagenesis process to proteotoxic stress. This is especially interesting in the light of the recent results from the REMoDEL-B study examining bortezomib as an addition to standard R-CHOP therapy in DLBCL. An unexpected outcome of a post-hoc analysis was that the addition of the proteasome inhibitor appeared to benefit patients of the MHG subtype. This study also proposes DDX3X as an attractive therapeutic target. DDX3Y is essential for maintaining the translational capacity of transformed B-cells, but is not

expressed at the protein level in normal cells (Ditton et al., 2004; Foresta et al., 2000a; Rauschendorf et al., 2011), which allows us to speculate that the toxicity of potential DDX3Y inhibitor will be low. Indeed, males with germline DDX3Y mutation present with infertility, but no other phenotype, except azoospermia, has been observed (Foresta et al., 2000a; Rauschendorf et al., 2011). Up to now, there are two small molecule inhibitors of DDX3 available for pre-clinical studies (Bol et al., 2015; Brai et al., 2016). Both target the core helicase domain of DDX3, which is almost identical between DDX3X and DDX3Y. Undoubtedly, there is an urgent need for the development of more DDX3Y-specific agents. An interesting solution could be the use of Proteolysis Targeting Chimeras (PROTAC) technology to target unique parts of DDX3Y to direct the protein to ubiquitin-proteasome system for degradation (Gao et al., 2020).

CHAPTER 5

Elucidating the role of translated micropeptides in Diffuse Large B-cell Lymphoma

5.1 Background

The human genome is thought to contain approximately 20,000 protein-coding sequences. The annotation of these Open Reading Frames (ORFs) has been based on a set of rules predicting translation of stable and functional proteins, for example requiring minimum length of 100 amino-acids (aa), methionine encoded translation start site, biased codon usage and high sequence conservation between species (Brent, 2005; Ingolia et al., 2011). The hypothesis of *one gene, one enzyme* (or one polypeptide in latter variants), formed initially in 1941 (Beadle and Tatum, 1941), also largely shaped our understanding of the organisation of the protein coding regions. However, in the light of recent studies, there are several exceptions from those rules suggesting that the human proteome might be much more complex and dynamic than we have thought.

Firstly, the number of variants and isoforms, or as has been recently proposed *proteoforms* (Smith and Kelleher, 2013), of known proteins may exceeds a fixed, gene-centric, set of 20,000 sequences. Some estimates put the figure of 70,000 or even few millions proteoforms produced in the human genome, if all post-translational modification are included (Aebersold et al., 2018a). Another layer of proteome complexity is provided by the discovery of new coding regions that were missed by the consensus annotations due to, for example, their small size or unexpected localisation. The existence of those noncanonical proteins, also referred to as *micropeptides* or *cryptic peptides*, have already been confirmed by several studies, which demonstrated wide range of biological activity of these products. For instance, the chemokine family, which are secreted by cells to

coordinate the immune response, comprise more than 40 known proteins with length between 67–127 amino-acids (Moser and Willimann, 2004). Other examples of biologically relevant micropeptides include the the family of defensins (18-45 aa), ribosomal protein L24 (RPL24, 25 aa) and three calcium transporter regulators: phospholamban (PLN, 52 aa), sarcolipin (SLN, 31 aa), and myoregulin (MLN, 46 aa). The ability of new sequencing techniques, such as Ribo-Seq to reveal the position of every translating ribosome in the cell opens new opportunities to re-examine the translation landscape. Ribo-Seq technique proved to be highly effective to identify new ORFs in *bona fide* non-coding regions of the genome, such as long non-coding RNAs (lncRNAs) or 5′ and 3′ untranslated regions, 5′UTR and 3′UTR respectively, flanking known ORFs (Chen et al., 2020; Chong et al., 2020; Jackson et al., 2018; Calviello and Ohler, 2017). However, only a handful of these have been validated to encode functional micropeptides with potent biological functions.

The scope of noncanonical translation and its exact role are still unclear. In attempt to bridge this gap, I took advantage of a large translatome dataset generated in the Hodson lab involving 79 Ribo-Seq libraries with matching RNA-Seq samples obtained from primary B-cells and lymphoma cell lines.

In this chapter:

1. I introduce a systematic approach to annotate *de novo* actively translated regions directly from Ribo-Seq data,

2. I query a set of mass-spectromentry datasets downloaded from public repositories for identified noncanonical ORFs,

3. I rank putative micropeptides according to their biological relevance,

4. I use the top scoring entries to the design a knockdown CRISPR screen to identify ORFs essential for B-cell growth and survival.

5.2 Results

5.2.1 A systematic approach for *de novo* identification of non-canonical translation products in lymphoid cells.

5.2.1.1 An integrated ORF identification workflow

To annotate regions of noncanonical translation, I developed a bioinformatic pipeline that integrates translatome (Ribo-Seq) and transcriptome (RNA-Seq) data creating a non-redundant database of putative peptides and proteins **(Figure 5.1 A)**.

The core idea of this workflow is to analyse the phasing pattern of ribosomal footprints that is expected to be enriched in actively translated regions. It is important to point out that a simple alignment of Ribo-Seq reads to a genomic region is not an evidence of active translation. Of RNAse protected mRNA fragments, sequenced in Ribo-Seq, only in about 85 % correspond to the ribosomes position. The remaining portion of footprints comes from other RNA-protein complexes (Ji et al., 2016). The two situations can be easily distinguished by looking at the distribution of Ribo-Seq reads over a queried region. Ribosome-derived reads (1) have length about 28-29 nucleotides, (2) span the entire translated region (signal uniformity), and (3) show 3-nucleotide periodicity in the signal shape (frame preference), which corresponds to the ribosome decoding only three nucleotides at time. In contrast, nonribosomal footprints are usually highly localised and produce reads of varying lengths (Ji et al., 2016).

To find genomic regions showing features of active translation I have combined four independent computational algorithms: ORFLine (Hu et al., 2021), ORF-RATER (Fields et al., 2015), RibORF (Ji, 2018) and RiboCode (Xiao et al., 2018) to **(Table 5.1)** that were developed specifically for this task. Overall, all four tools benefit from a regular pattern of alignment of ribosome-derived footprints, but adopt different statistical models to rank and select regions, which have the highest probability of being translated. Some tools incorporate also additional metrics that reflect signal uniformity, such as Inside/outside ratio (I/O ratio), sequence coverage or Ribosome Release Score (RRS) quantifying the abrupt drop in Ribo-Seq signal after the stop codon.

Because translation can sometimes begin from codons other than AUG (Kearse and Wilusz, 2017), all possible reading frames constrained by any combination of start (AUG, UUG, GUG or CUG) and stop codons (UAA, UGA, or UAG) were considered. All parameters of ORF-finding algorithms were kept at default values, as recommended in the original publication, see section 2.2.11 for details.

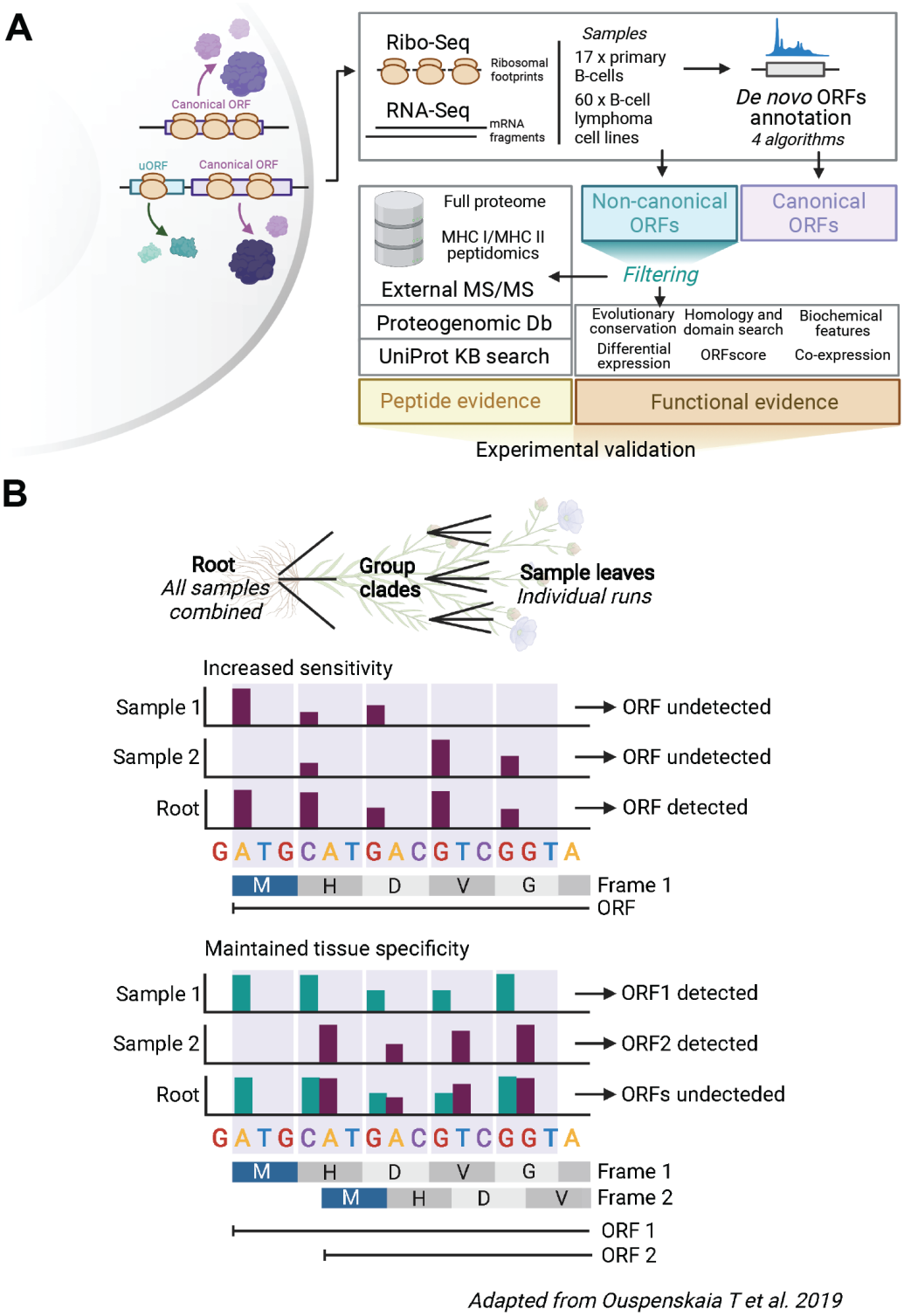

Figure 5.1: Systematic approach to identify translated ORF in lymphoid cells

A) Flow chart showing the workflow for identifying ORFs from Ribo-Seq datasets

B) Diagram explaining the strategy of hierarchical merging adapted from Ouspenskaia et al. (2020). Samples are represented as *leaves* that are merged based on the biological similarity to create *clades* and finally one large *root* file containing all sequenced reads. Accumulation of Ribo-Seq reads over multiple samples should increase the strength of frame preference (increase sensitivity), while maintaining tissue specificity, as individual samples are also used for ORF identification.

Table 5.1: An Overview of ORF-finding algorithms used in this study

Tool	Input	Start codons	Translation metrics	Statistical evaluation
ORFLine	Ribo-Seq RNA-Seq	NUG	Frame preference, Ribosome release score, Inside/outside read ratio, sequence coverage	Log-scaled chi-squared goodness of fit test for frame preference
ORF-RATER	Ribo-Seq RNA-Seq	NUG	Frame preference, aggregate profiles over start/stop codons	Linear regression model
RibORF	Ribo-Seq	NUG	Frame preference, signal uniformity	Logistic regression model Support Vector Machine classifier
RiboCode	Ribo-Seq	NUG	Frame preference	Modified Wilcoxon signed-rank test

First, the RiboStream pipeline was applied to a large dataset of 79 paired Ribo-Seq and RNA-Seq libraries covering a variety of lymphoid cell types including 12 samples from primary germinal centre B-cells, 5 samples from DLBCL tumour biopsies and 60 samples from established lymphoma cell lines.

To maximise sensitivity and specificity of detection of translated regions, I adopted a strategy of hierarchical ORF detection that was initially introduced by Ouspenskaia et al. (2020). If the number of aligned footprints fall below certain level, the 3-nucleotide periodicity become indistinguishable from non-translational noise and may be missed by ORF identification algorithm.

However, if we aggregate the Ribo-Seq signal across multiple samples and run the tools on the merged set of footprints, truly translated regions should accumulate enough mapped reads to show a clear phasing pattern, thus, reaching detection level of ORF finding algorithms. Hierarchical merging of Ribo-Seq and RNA-Seq BAM files (aligned reads) is guided by the biological similarity, which allows to increase sensitivity of detection while maintaining the specificity of cell type expression **(Figure 5.1 B)**. For example, from the data generated in chapter 3 (3 experimental conditions, 4 replicates each) 4 merged files can be built and 16 ORF identification runs can be performed at three levels:

1. Leaves level: individual samples runs (12 samples)
2. Group (clades) level: files merged by experimental condition (3 files, one per condition))
3. Root level: all 12 individual samples merged into 1 file

The same logic has been applied to all 79 Ribo-Seq samples giving rise to 177 files to run ORF identification tools.

5.2.1.2 Pervasive translation of crude non-coding regions in lymphoid cells

In total, I found 13 483 canonical ORFs and 43 537 noncanonical ORFs identified by at least two algorithms **(Figure 5.2 A)**. Interestingly, only 420 (0.7%, 420/57,020) noncanonical ORFs were picked by all four tools what suggests substantial differences between the ORF finding strategies and reinforces the need of using more than one tool for comprehensive identification. As expected, the number of ORFs was positively correlated with the number of ribosomal footprints used as an input. The merged files, generated through hierarchical merging strategy, brought almost one order of magnitude more ORFs than would be from using from just *raw*, unmerged, files **(Figure 5.2 B)**.

Although all tools classify identified ORFs into distinct types routinely, the strategies to do so and the exact definitions of those ORF types vary between the algorithms. To make the results comparable, I reassigned the ORFs to a representative transcript isoform and unified ORF classification specifying nine ORF categories based on their genomic location and reading frame **(Figure 5.2 C)**, see section 2.2.11 for detailed description of the classification strategy.

Of the 43 537 noncanonical ORFs, the largest group (31232/43537, 71.8%) encompassed ORFs situated in the noncoding parts of coding transcripts, or overlapping the known CDS out-of-frame with the annotated product. The noncanonical ORFs were found in 8586 protein-coding transcripts and each of these contained more than one additional ORF, 2.8 on average. 14.7% (6386) of noncanonical ORFs showed substantial, in-frame, overlap with annotated ORFs and were considered new isoforms of known ORFs with either truncated or extended amino-acid sequence. Among the noncanonical ORFs, 5919 (5919/43537, 13.6%) were localised on transcripts annotated as noncoding, predominantly in long noncoding RNAs **(Figure 5.2 C)**.

Consistent with previous studies (Chen et al., 2020; Cuevas et al., 2021), noncanonical ORFs were much shorter than the canonical proteins: the length of almost 50 % was lower than 100 amino-acids **(Figure 5.2 D)**. Noncanonical ORFs showed strong frame preference, accumulation of footprints at the start codon and sudden drop in signal after estimated termination site **(Figure 5.2 E)**, which all characterise active translation and resembles the pattern observed in known protein coding regions.

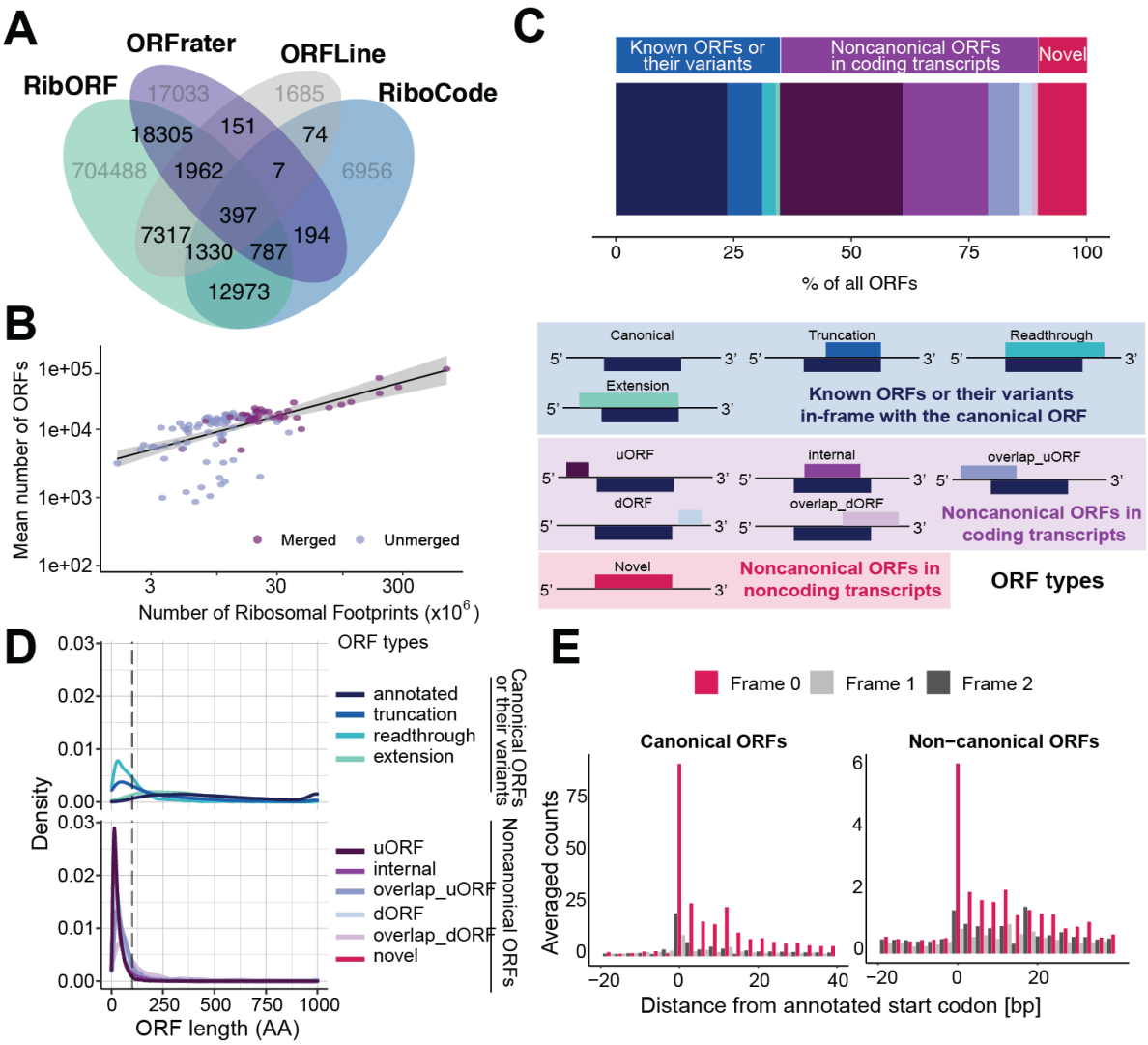

Figure 5.2: Ribo-Seq reveals pervasive translation of noncanonical ORFs in B-cells.

A) Venn Diagram showing the total number of ORFs identified and the overlap between the ORF-finding algorithms. Only ORFs detected by more than 1 tool (numbers in black) were selected for further analysis.

B) Scatter plot showing the relationship between the mean number of ORFs detected per sample and the total number of ribosomal footprints used as an input. The colour indicates which input files were obtained from merging the original FASTQ files together.

C) Barplot showing the percentage contribution of each ORF type to the total of 57 020 ORFs identified. ORF types were divided into three classes: known ORFs and their variants (blue), noncanonical ORFs localised on coding transcripts (violet) and novel ORFs in noncoding transcripts (pink).

D) Histogram showing ORF length distribution for canonical and noncanonical ORFs.

E) Barplot showing average three nucleotide periodicity and single frame preference in canonical and noncanonical ORFs.

5.2.2 Noncanonical ORFs account for about 10% of proteins detected in proteomics experiments

The analysis of the Ribo-Seq dataset revealed pervasive translation of thousands of noncanonical regions in lymphoid cells. The most burning questions arising in this context are: 1) are those products of noncanonical translation detectable at the protein level, 2) is this a reproducible finding, and 3) are there any additional evidence, other than Ribo-Seq, that those regions are biologically relevant?

I addressed the first question by reanalysing 13 publicly available mass spectometry (MS) datasets, covering various cell types and conditions, and querying them for the presence of peptides originating from predicted ORFs. Searching for the products of noncanonical translation in proteomic data is a nontrivial challenge. Firstly, their small molecular weight can impede detection by standard mass spectrometry techniques, thus, even less popular methods should be explored. Recent studies demonstrated that noncanonical proteins are enriched in Major Histocompatibility Complex (MHC) bound peptides (MAPs) (Chong et al., 2020; Chen et al., 2020; Ouspenskaia et al., 2020). Mass-spectrometry based identification of MAPs (immunopeptidomics) has attracted much attention in the context of tumour neoantigens and numerous datasets are deposited in public repositories. Another interesting technique is, so called, *deep proteome* or high-resolution MS, which allows for higher protein coverage and broader dynamic range of detection (Bekker-Jensen et al., 2017). Therefore the collected datasets included SILAC or tandem mass tag (TMT) techniques, as well as immunopeptidomics and deep-proteome experiments, see section 2.2.12 for details regarding the bioinformatic workflow.

A typical analysis of a proteomic dataset involves the assignment of collected spectra to the reference sequence. This poses another difficulty, because an inflated search database used for mass spectra survey can lead to spurious matches increasing the number of false positive findings (Blakeley et al., 2012; Nesvizhskii, 2014). To decrease the search space for MS analysis 1) I narrowed down a list of putative ORFs by removing all ORFs shorter than 18 nucleotides (6 amino-acids), and 2) filtered out all in-frame ORFs with more than 20% overlap with known ORFs focusing my attention on unique ORFs that are unlikely to be variants of known peptides. After filtering my list of putative micropeptides had 30,188 unique sequences that were used to build a customised search database for MS analysis. To avoid forced assignment of the mass spectra to noncanonical ORFs, the final search database contained the noncanonical peptides sequences and a set of reference proteins downloaded from UniProtKB. I queried the MS datasets using MaxQuant software or NewAnce workflow, designed specifically to process the immunopeptidomics data, see Chong et al. (2020).

From all MS datasets surveyed, I identified 564,764 peptides assigned to 12,518 unique proteins (PSM FDR 1%, Protein FDR 1%) **(Figure 5.3 A)**. Of these, between 1 and 5.5 % were assigned to noncanonical ORFs **(Figure 5.3 B)** accounting for 1,311 unique ORFs in total. The total number of peptides identified in MS data and the percentage of those derived from noncanonical ORFs varied slightly between datasets and MS techniques, the difference was on the border of statistical significance (Kruskall-Wallis $p = 0.1$). The highest proportion of peptides derived from noncanonical ORFs was in MHC-I immunopeptidomics experiments and in a dataset utilising a *deep proteome* technique (Bekker-Jensen et al., 2017) **(Figure 5.3 B)**. This may suggest that certain properties of immunopeptidomics and deep proteome experiments may facilitate the detection of noncanonical proteins. It can be attributed to technical as well as biological properties, such as preferential access to the antigen presentation pathway, difficulties with detecting small or short living proteins by standard techniques. MHC complex pull-down, multi-fractionation or the usage of multiple proteases in deep-proteome studies may facilitate detection of otherwise *undetectable* proteins (Bekker-Jensen et al., 2017).

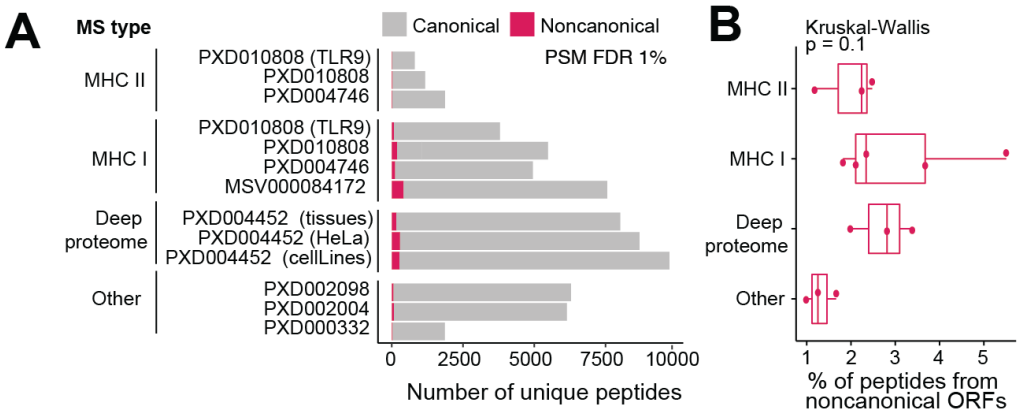

Figure 5.3: Analysis of mass spectrometry datasets in search of peptides matching predicted noncanonical ORFs

A) Barplot showing the total number of unique peptides identified in each mass spectrometry dataset. Colour of the bar indicates whether a peptide has been assigned to noncanonical (red) or canonical (grey) protein.

B) Boxplot showing the percentage of unique peptides matching predicted noncanonical proteins. Statistical significance of the difference in the proportion of peptides derived from noncanonical ORFs in each MS group was determined with Kruskal-Wallis test, p-value = 0.1.

Next, I evaluated the accuracy of noncanonical proteins identification using four metrics. In all datasets the distribution of mass measurement error (the difference between an individual measurement and an expected value for a peptide) and the Andromeda score (probabilistic score reflecting the accuracy of peptide to MS spectrum match) were indistinguishable for noncanonical and canonical proteins **(Figure 5.4 A-C)**. The observed Retention Time (RT) of eluted peptides was highly correlated with the chromatographic Hydrophobicity Index (HI) predicted from the amino-acid composition of a peptide **(Figure 5.4 C)** (Krokhin and Spicer, 2010). The correlation for peptides derived from canonical and noncanonical ORFs was similar. Lastly, the distribution of MHC-bound peptides was similar in both groups with the majority of peptides having length between 9 and 10 nucleotides for MHC-I and 11-17 nucleotides for MHC-II **(Figure 5.4 E)**.

Finally, I addressed the question of reproducibility of ORF identification. I aligned the noncanonical ORFs against the extended database of human proteins downloaded from UniProt (SwissProt and TrEMBL sets) and from two public repositories, OpenProt and sORFdb, containing the sequences of new proteins predicted by other proteogenomic studies. 5774 (19.12%, 5774/30188) ORFs matched UniProt, OpenProt or sORFdb with at least 95% sequence similarity **(Figure 5.4 F)**. Almost half of the ORFs (47.52%, 623/1311) with matching peptides from the reanalysed MS experiments were confirmed in at least one of the databases **(Figure 5.4 F)**.

In summary, although relatively small proportion of noncanonical ORFs was identified in proteomic experiments, they accounted for about 10% of all unique proteins quantified and showed the accuracy of detection no different to known proteins. In total, 21.40% of noncanonical ORFs (6462/30188) had evidence of protein levels expression (UniProt or proteomics) or have been identified as actively translated by other groups, which reinforces the authenticity of our Ribo-Seq workflow. The remaining 78.6% were unique to Ribo-Seq data from this study (no protein level evidence and not identified in other translatome studies).

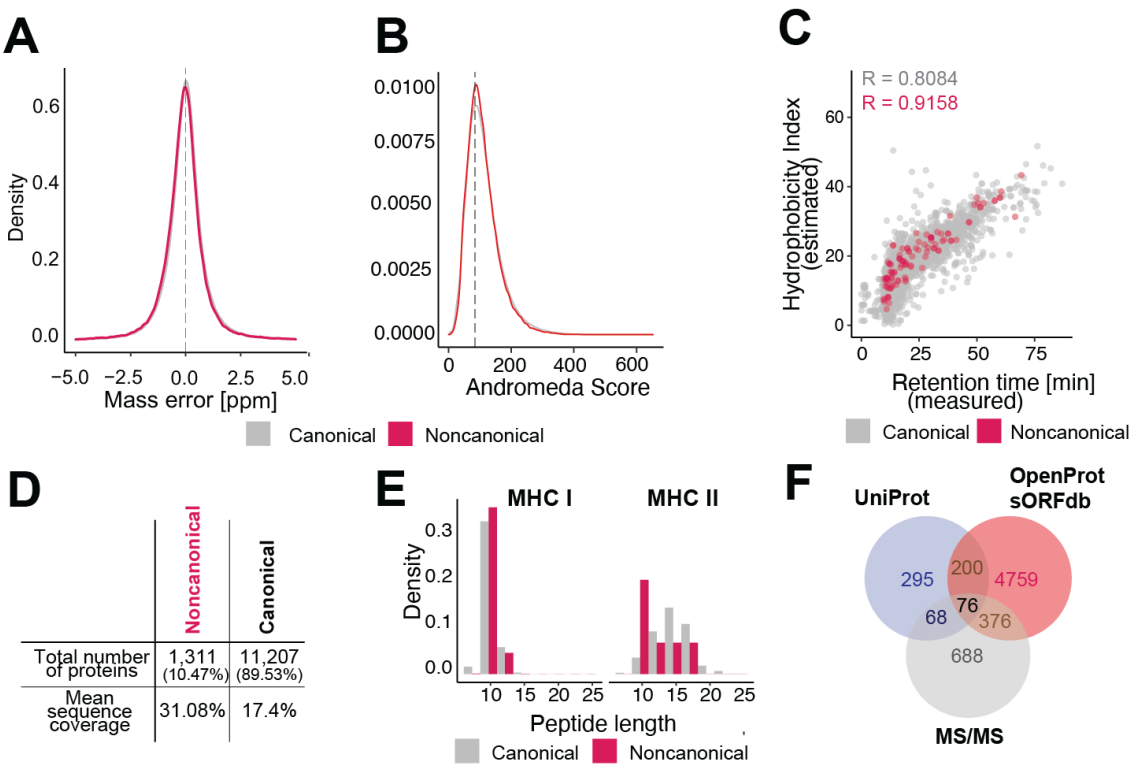

Figure 5.4: Features of peptides from canonical and noncanonical proteins identified in MS experiments

A) Histogram of mass error for identified canonical and noncanonical peptide-spectrum matches (PSM). The p-value $< 10^{-16}$, calculated with Kolmogorov–Smirnov test (Number of canonical PSM = 7378711, number of noncanonical PSM = 80829).

B) Histogram of Andromeda Score, which mirrors the accuracy of peptide-spectrum match of identified canonical and noncanonical peptides. The p-value $< 10^{-16}$, calculated with Kolmogorov–Smirnov test (Number of canonical PSM = 7378711, number of noncanonical PSM = 80829).

C) Pearson correlations between observed and SSRCalc predicted retention times of peptides derived from canonical and noncanonical proteins.

D) Total number of unique canonical and noncanonical proteins and mean sequence coverage of matched peptides.

E) Histogram of length of MAPs derived from canonical and noncanonical proteins.

F) Venn Diagram showing the overlap between ORFs identified in MS datasets, found in UniProt and proteogenomic databases.

5.2.3 Characteristics of noncanonical ORFs producing MHC-bound peptides

Next, I wished to find the features of noncanonical proteins that increase the probability of protein level expression. In contrast to known proteins, there was only a little overlap in noncanonical ORFs detected in immunopeptidomics and full proteome MS data: only 4.3% were identified by both techniques **(Figure 5.5 A-B)**. The distribution of the ORF types between various proteomic techniques was also remarkably different **(Figure 5.5 C)**. MHC-bound peptides were more likely to come from uORF, while peptides derived from noncoding transcripts (pseudogenes or lncRNAs) were more common in full proteome data (Chi-square test for independence, p-value $< 2.2 \cdot 10^{-6}$). Interestingly, there was no difference in the distribution of ORF types between *standard* MS experiments and deep-proteome studies (Chi-square test for independence, p-value = 0.59). This suggests that there might be a relationship between the ability to detect ORF-derived peptides and their biological properties.

I integrated different ORFs characteristics into a machine-learning algorithm based on Random Forests to predict MS detection. Because of little overlap between immunopeptidomics and full proteome studies in detecting peptides originating from noncanonical ORFs, I built separate single-class classification models to distinguish MHC-detected vs. not detected and full proteome (deep proteome or standard MS) detected vs. not detected. Sequence characteristics incorporated into the models included: the percentage of the overlap with known CDS, ribosomal footprints density, translation efficiency, gene type (protein-coding, lncRNA or pseudogene), GC content, length, amino-acid composition and isoelectric point (pI). Variants or isoforms of known genes (ORFs classified as canonical, truncated, readthrough or extended) were filtered out and all noncanonical ORFs with mean mRNA expression and ribosome footprints density > 0 were included. Of these randomly selected 75% were used as a training set, the remaining 25% was used for testing. The performance varied substantially between the two suggesting that ORF features are not always sufficient to predict detection in proteomics. ORF characteristics predicted well the identification of MHC-bound peptides (80% recall at 40% precision, AUPRC = 0.4919). Chromosome location, ORF score and the number of ORF identifying tools recognising an ORF as translated were the most important factors for accurate prediction. ORFs with detected MHC-bound peptides were significantly enriched in chromosome 4, 20 and 6 (z-test of two proportions, adjusted p-value < 0.01), had higher ORF scores (Kolmogorov-Smirnov test, p-value $< 10^{-16}$) and were more likely to be detected by more than one tool (Chi-squared test $< 10^{-16}$). Surprisingly, despite similar number of ORFs detected in full proteome studies, the prediction did not work efficiently for this group (AUPRC = 0.0264).

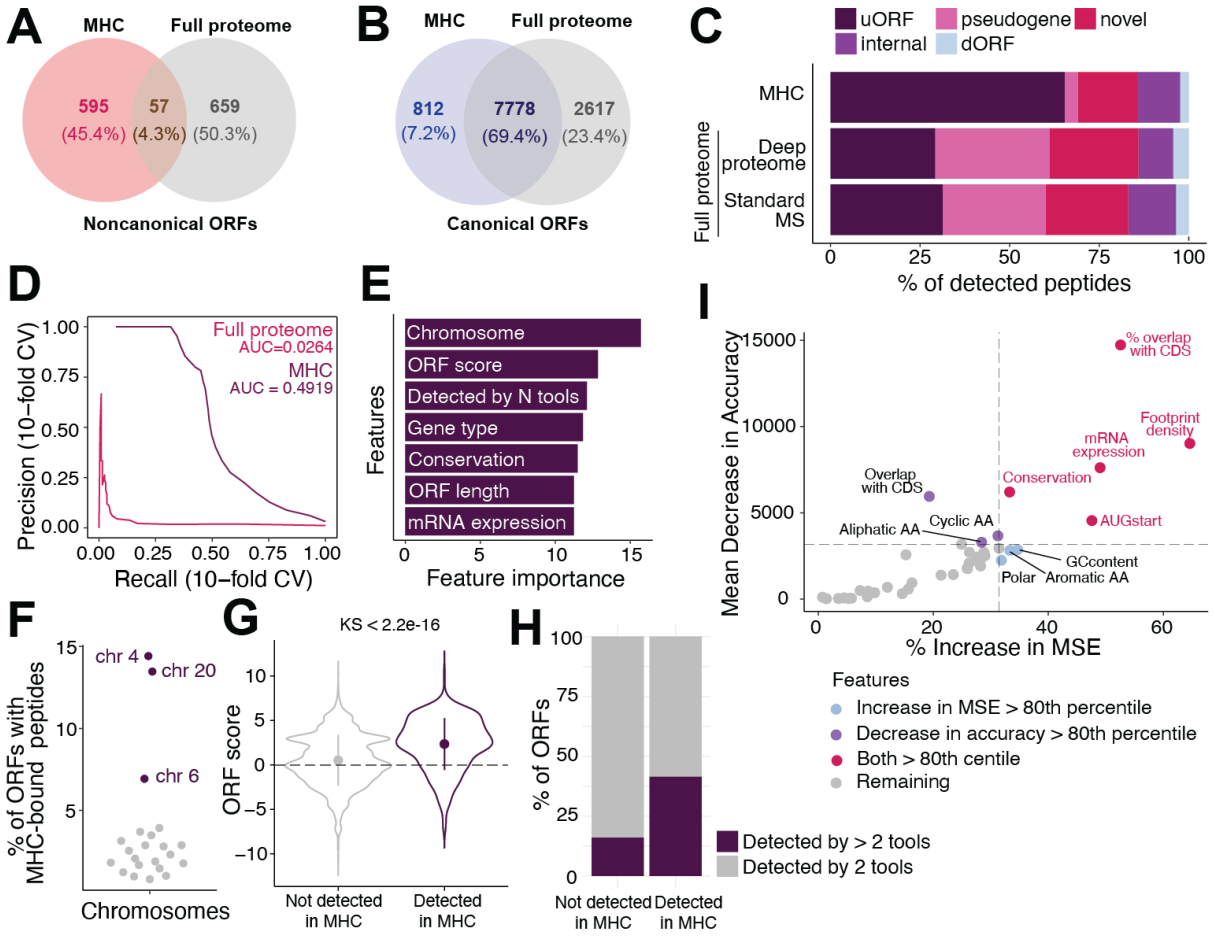

Figure 5.5: Ribo-Seq reveals pervasive translation of noncanonical ORFs in B-cells.

A) Venn Diagram showing the overlap between MHC and full proteome derived peptides derived from noncanonical proteins.

B) Venn Diagram showing the overlap between MHC and full proteome derived peptides derived from canonical proteins.

C) Barplot showing the proportion of ORF types in peptides identified by different mass spectrometry techniques. Chi-square test for independence, p-value $< 2.2 \cdot 10^{-6}$

D) Performance of machine-learning-based classifiers in predicting detection of noncanonical ORFs derived peptides in mass spectrometry. Random forest classifiers were trained on the set of noncanonical ORFs characteristics and performance was assessed in a tenfold cross-validation (CV) mode.

E) Feature importance of random forest classifier predicting MAPs detection represented as mean decrease in accuracy.

F) Percentage of noncanonical ORFs with MAPs per chromosome. Highlighted points indicate chromosomes with statistically significant enrichment of MAPs. Z-test of two proportions, adjusted p-value < 0.01.

G) Violin plot showing the difference in ORF score between noncanonical ORFs with and without detected MAPs. Kolmogorov-Smirnov test, p-value $< 10^{-16}$

H) Percentage of noncanonical ORFs detected with more than one tool compared for ORFs with without detected MAPs, Chi-squared test for independence $< 10^{-16}$

I) Scatter plot showing feature importance of random forest regression model predicting the value of the ORF score; measured with mean decrease in accuracy and mean increase in mean squared error (MSE).

I also investigated the factors determining ORF score value, which directly translates into the strength of three-nucleotide periodicity, and so is the key determinant of active translation. Overall ORF features explained 61.29 % of ORF score variance with overlap with known CDS, ribosome footprint density, mRNA expression, conservation score and AUG as a start codon being the most influential predictive features. High importance of expression related measures (mRNA expression or ribosomal footprints density) resonates well with known relationship between the number of ORFs annotated and the number of mapped ribosomal footprints.

This analysis suggests that, firstly, certain biological factors, sequence characteristics may determine MHC presentation of peptides derived from noncanonical ORFs and, secondly, low sequence coverage or low expression level may be the primary limiting factor for efficient detection and validation of noncanonical ORFs.

5.2.4 Design of customised knockout CRISPR screen to identify noncanonical ORFs important for B-cells survival

To identify the noncanonical ORFs that are essential for B-cell survival in a systematised manner, I wished to screen the most promising candidates using a customised CRISPR library which contains 6000 gRNAs targeting 1,625 not overlapping ORFs.

I designed a knockout CRISPR screen to target selected ORFs with features indicating an important biological function in lymphoid cells. I divided the process of selecting ORFs for screening into two stages: negative selection aiming to discard all ORFs that are not fit for CRISPR-Cas9 targeting and positive selection enriching for regions with interesting biological features.

The first stage discards all ORFs which are either too short to be targeted with sufficient number of gRNAs, have low expression in the primary cells or cell lines chosen for the screen, or share too large overlap with known CDS region, so that the observed phenotype could be easily explained by disruption in the canonical coding region **(Figure 5.6 A-B)**. Next I screened the remaining ORFs for compatible gRNAs cutting within a predicted ORF. I filtered out all *untargetable* ORFs with the number of good quality gRNAs below 5. By *good quality* I understand gRNAs with GC content between 40 and 70 %, with no direct off-target locations and without homopolymers of 4 or more consecutive Ts. The presence of homopolymers in gRNA sequence can decrease cutting activity and TTTT is known to act as minimal T-stretch termination signal for RNA polymerase III (Gao et al., 2018). Indeed, when I compared the efficiency of dropout for genes identified as essential in knockout CRISPR-Cas9 screen in lymphoid cell lines (Phelan et al., 2018), I observed significant decrease for gRNAs containing homopolymers, especially series of Ts **(Figure 5.6 C)**.

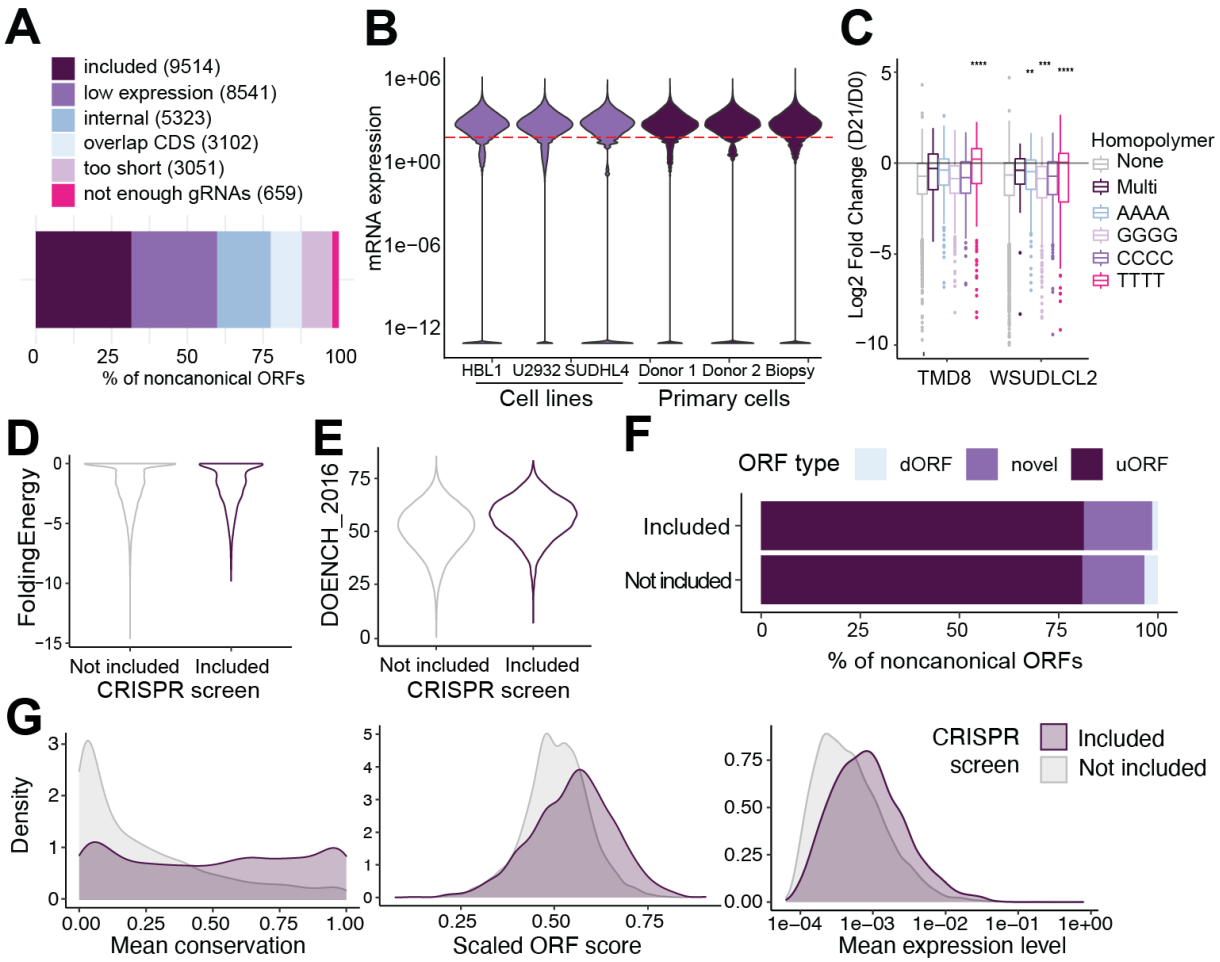

Figure 5.6: Design of a CRISPR-Cas9 CRISPR screen library to study noncanonical ORFs in lymphoid cells

A) Barplot showing the proportion of noncanonical ORF excluded from the screen. In total about 68.49 % ORFs were excluded because of low expression level, too large overlap with known CDS or too being too short.

B) Violin plots showing mean expression level of noncanonical ORFs in cells selected for the screen. Red dashed line corresponds to the 20th percentile, which was considered a threshold for low expression.

C) Barplots showing dropout efficiency of gRNAs targeting essential genes from Phelan et al. (2018) knockout CRISPR-Cas9 screen stratified by the presence of homopolymers in gRNAs sequence. Statistical significance (compared to 'None' group) determined with two sample Wilcoxon test, adjusted p-values computed with Benjamini and Hochberg method. **** $<$ 0.0001, *** $<$ 0.001, ** $<$ 0.01

D) Violin plot showing RNA folding energy of gRNA included and not included in the final CRISPR-Cas9 screen library.

E) Violin plot showing one of the gRNA efficiency scores for gRNAs included and not included in the final CRISPR-Cas9 screen library.

F) Barplot showing the percentage of different ORF types targeted with the library of gRNAs.

G) Histograms showing the ORFs targeted with the gRNA library have overall higher evolutionary conservation, higher ORF score and higher expression level.

The p-values in all groups (folding energy, DOENCH 2016 score, evolutionary conservation, ORF score and expression level) $< 10^{-10}$, calculated with Kolmogorov–Smirnov test

The efficiency of all gRNAs was evaluated using a number of efficiency scores, including prioritising a G at position 20 (upstream of PAM) and few multi-factor scores predicting gRNA stability and activity (Moreno-Mateos et al., 2015; Xu et al., 2015; Doench et al., 2014, 2016) **(Figure 5.6 D)**. Given reports about inferior on-target activity of gRNAs with internal harpins and regions of self-complementarity, I also prioritised gRNAs with lower folding energy (Thyme et al., 2016) **(Figure 5.6 E)**.

9514 ORFs that left after the stage of negative selection were then ranked based on the features that make them interesting from a biological perspective. I took intro account the following characteristics: mean ORF score (reflecting the strength of frame preference), mRNA expression level, the fact of being detected at the protein level (in immunopeptidomics or full proteome data) or matching predicted proteins from external databases, evolutionary conservation, the number of tools that identified an ORF as actively translated, the number of samples where this happened and the differential expression pattern shown in our large Ribo-Seq dataset. The final library targets 1,625 top scoring ORFs with gRNAs showing the highest predicted on-target activity, on average 4-5 gRNAs per ORF **(Figure 5.6 F-G)**.

As negative controls I used 250 non-targeting gRNAs from an established Brunello CRISPR-Cas9 library (Sanson et al., 2018), which do not recognise any sequence in the human genome. This will help me to distinguish the effect of neutral drift from selective disadvantage caused by disruption of biologically important ORFs. For positive control I selected two known oncogenes, MYC and POU2AF1, which knock-out should be disadvantageous for established lymphoma cell lines and primary GC B-cells (Phelan et al., 2018; Caeser et al., 2019) allowing me to estimate the expected dropout capacity of the screen. This library was ordered as an oligo pool from Twist Bioscience and cloned into a lentiviral backbone, maintaining representation $>$ 100 colonies per guide. It was then introduced into six Cas9-expressing lymphoma cell lines and ex vivo human GC B cells maintaining representation of $>$1,000 transduced cells per guide. Cells were harvested at day 21 and sequencing libraries prepared. At the time of writing these libraries are waiting to be sequencing. These CRISPR screening experiments were conducted by Dr. Stamatia Vori, a Masters student in the Hodson lab.

5.3 Discussion

The last four years have witnessed an increase in studies investigating the topic of noncanonical translation. Our extensive collection of 79 B-cell translatomes generated in primary GC and malignant B-cell cells provides a powerful resource to investigate this question in the context of physiology and pathology. To our knowledge, this is the largest study exploring this topic in B-cells so far. The proteogenomic method described here integrates Ribo-Seq, RNA-Seq, mass spectrometry (MS) and external databases to comprehensively evaluate the scope of noncanonical translation in lymphoid cells and identify the most promising targets for high-throughput knockout CRISPR screen. The initial results of the systematic analysis of B-cells translatomes revealed 43 537 noncanonical Open Reading Frames (ORFs) and 13 483 canonical (known) ORFs. Noncanonical ORFs were typically situated in ostensibly noncoding regions of the genome, accounted for about 10% of all proteins detected in the analysis of external mass spectrometry experiments, and almost 20% was also found in other proteogenomic databases. To identify noncanonical ORFs essential to B-cell survival, I designed a customised knockdown CRISPR screen library targeting top ORFs with the highest likelihood of protein level expression and biological relevance.

Evidence behind pervasive translation of noncanonical ORFs

This and previous studies (Chen et al., 2020; Chong et al., 2020; Cuevas et al., 2021; van Heesch et al., 2019) employed Ribo-Seq to provide evidence for widespread ribosome occupancy in noncoding regions of the transcriptome. This has been observed in a range of model organisms (Zhang et al., 2019; Mackowiak et al., 2015), in studies utilising different translation inhibitors for Ribo-Seq library preparation (Ingolia et al., 2011; Zhang et al., 2018a; Lee et al., 2012), in polysome-centric studies (Poly-Ribo-Seq) (Aspden et al., 2014), and *in vivo* with RiboTag RNA sequencing (Jackson et al., 2018; Sanz et al., 2009). This is a striking finding that puts some rules of eukaryotic translation into question, including monocistronic organisation of eukaryotic transcripts or the purely noncoding nature of several long-noncoding RNAs.

Limitations of proteomic-based approaches for novel proteins discovery

Although noncanonical translation seems to be a recurrent observation in translatome studies, validation of their products at the protein level is lagging behind. In a most popular approach - shotgun proteomics, proteins extracted from the cell are fragmented with a protease (typically trypsin), and the mixture of digested peptides is analysed with tandem mass spectrometry (MS/MS). Identification of the mass spectra occurs through comparison with the reference database of *in silico* generated spectra from a provided

sequence of reference proteins (Perez-Riverol et al., 2018). This poses difficulties for pinpointing peptides originating from proteins not included in the reference sequence database, with unexpected post-translational modifications (PTMs) or with low signal-to-noise ratio (Griss et al., 2016). On the other hand, an inflated reference database leads to a high number of false positive findings (Blakeley et al., 2012; Nesvizhskii, 2014), so the reference must be prepared with care. Here, I adopted a parsimonious strategy combining the sequence of known proteins with the sequence of ORFs predicted with Ribo-Seq, so that using the smallest possible reference database to search for novel proteins.

Only about 4% of noncanonical ORFs predicted in our study had evidence of protein level expression. Are those remaining 96% of noncanonical ORFs a technical noise, or we just failed to detect it with proteomics? Both scenarios are possible. The arguments in favour of the first are: relatively low reproducibility of the assay between different ORF identification algorithms and the small overlap with external proteogenomic databases as only about 20% of noncanonical ORFs were observed in other studies. Another issue is that we cannot exclude the possibility that the Ribo-Seq signal, even when showing the periodical pattern of alignment, does not always correspond to actively translating ribosomes. For example, could the non-coding regions of the genome work as a *ribosome sponge* by sequestrating them or directing their transport? Noncanonical translation could also correspond to mRNA quality check or maturation process, e.g. during a pioneering round of translation (Maquat et al., 2010).

On the other hand, although orthogonal evidence for about 20% of noncanonical ORFs seems like a low recall, it may be a lot given that we know very little about noncanonical translation biology. Even in the group of 20,386 manually curated human proteins (Swiss-Prot), almost 20% (3,989/20,386) lacks strong protein level evidence (Perdigão et al., 2015). Indeed, out of about 20,000 canonical proteins, only half was identified in our MS data analysis. On average, 75% of mass spectra reported in a typical MS experiment remains unidentified (Griss et al., 2016). Many of these have high quality and are likely to emerge from real proteins (Chick et al., 2015). The life cycle (synthesis and degradation rate), a pattern of PTM or subcellular localisation of noncanonical ORFs may be different to canonical proteins, which may poses difficulties with detection using standard proteomic techniques. For example, not all noncanonical micropeptides contain a tryptic cleavage site, which may complicate their identification in assays using trypsin for protein fragmentation. Enrichment of noncanonical translation products in the MHC-bound peptidome (immunopeptidome) is an unprecedented finding (Chong et al., 2020; Ouspenskaia et al., 2020; Chen et al., 2020; Cuevas et al., 2021). Firstly, it reinforces the concept that the search for micropeptides, 'hidden' in the cellular proteome, should include a broad range of proteomic techniques. And secondly, it suggests that the role of micropeptides may be linked to immunity and the generation of peptides directed for

MHC presentation. Finally, the MS data analysed here were downloaded from external repositories, so there is a possibility that our ability to detect mass spectra matching predicted micropeptides might be lower than in studies, where MS and Ribo-Seq data were generated from the same biological model. For example, matching mass spectra were found for about 7% of noncanonical ORF identified by Chen et al. (2020).

Potential role of micropeptides in immunity and immune surveillance

Noncanonical translation has drawn recently much attention in the context of antigen presentation, and prospective targets for cancer immunotherapy as a considerable fraction of MHC-derived peptides (MAPs) has been attributed to noncanonical proteins (Chong et al., 2020; Ouspenskaia et al., 2020; Cuevas et al., 2021). In line with those studies, I also observed a sizeable proportion of MAPs (between 2 and 5%) originating from noncanonical ORFs. Of 9,242 proteins identified in all reanalysed immunopeptidomics datasets, 652 (7.05%) were noncanonical proteins encoded by ORFs from noncoding regions of the genome. MHC genes are one of the most polymorphic genes, and each allele (allomorphs) binds a distinct set of peptides. This may limit our ability to identify MAPs accurately if the immunopeptidomics data were generated from different cells than the Ribo-Seq samples used for ORF prediction.

A fascinating concept linking noncanonical translation to immunopeptidome and defective ribosomal products (DRiPs) was formulated by Jonathan Yewdell. DRiPs are short-lived peptides (half-lives of minutes), which originate from invalid rounds of translation, e.g. due to mutations, synthesis errors, misfolding, truncation etc. (Dersh et al., 2021). Initially, the DRiPs hypothesis was put forward to explain the rapid presentation of antigens derived from stable viral proteins (Yewdell et al., 1996). DRiPs can also arise in abundance from noncanonical translation (e.g. ORFs with near-cognate start codons), especially when 5′-cap-dependent translation is shut down during stress or viral infection. Interestingly, the composition of MAPs is poorly reflected by the transcriptome and proteome, i.e. most abundant mRNAs or proteins do not necessarily produce the largest number of MAPs (Pearson et al., 2016). Certain genomic regions are 'hot spots' of MAPs despite relatively small contribution to the cellular proteome or transcriptome (Pearson et al., 2016). Preferential access of certain peptide groups to the MHC presentation pathway could be a part of the immunosurveillance process, which is especially important in the context of tumour formation. Immunosurveillance involves a complex interplay between the tumour immunogenicity, immune cells infiltration, cytotoxic T-cells activation, immune checkpoints and the microenvironment (Dersh et al., 2021). A better understanding of cancer-specific antigens could aid in designing cancer vaccines, personalised CAR T cell therapy or drugs that increase peptide generation in cancer cells, thus immune *visibility* of the tumour.

Possible roles of upstream Open Reading Frame (uORFs)

Translation of thousands of upstream Open Reading Frames (uORFs) is a recurrent finding in global translatome studies, including this one. The regulatory properties of uORFs have been known for a long time, but the evidence was mainly anecdotal, limited to a handful of transcripts, such as ATF4, MDM2, CEBPA and CEBPB (Wethmar et al., 2014). However, current estimates place the percentage of mammalian protein-coding genes with potentially functional uORFs at around 40-50% (Johnstone et al., 2016; Lee et al., 2012). In our study, I identified uORFs in 37% (5034/13483) of protein-coding genes, between 2 - 3 uORFs per gene.

uORFs are usually permissive for translation of the main coding sequence, but the efficiency of re-initiation of the downstream translation may be reduced Smith et al. (2021); Zhang et al. (2019); Hinnebusch et al. (2016). Despite a possible deleterious effect of uORFs on canonical ORF translation and observed depletion of population variants creating polymorphic uORFs, uORFs present in the genome show a higher level of conservation than expected from neutral evolution (Zhang et al., 2019; Churbanov et al., 2005; Zhang et al., 2021). A recent systematic analysis of 16,907,129 upstream AUGs (uAUGs) in 478 eukaryotic species showed strong purifying selection for the vast majority of uORFs suggesting the biological importance of these regions (Zhang et al., 2021). In line with these studies, I observed many uORFs localised in the regions of high evolutionary conservation.

Up to now, regulation of translation of the downstream (canonical) ORF is the main role of uORF. The amino-acid sequence of putative, uORF-encoded micropeptides shows a smaller degree of conservation, suggesting that the primary function of the majority of uORFs might be a fine tuning of downstream ORF translation rather than the encoding of stable proteins (Zhang et al., 2021). The balance between uORF and main ORF translation is influenced by a number of factors. Proposed determinants include: the total number of uORFs in 5′UTR (Zhang et al., 2018a; Chew et al., 2016), uORF position with respect to canonical start codon (distance, out-of-frame/in-frame) (Johnstone et al., 2016; Chew et al., 2016; Calvo et al., 2009), the activity of certain translation factors, Kozak sequence context (Rogozin et al., 2001) or the adjacent mRNA secondary structure (Chew et al., 2016; Zhang et al., 2019).

A well studied mechanism of preferential uORFs translation is stress and immune response. The frequency of uORF *leaky scanning* (start codon skipping) negatively correlates with the availability of the ternary complex (Orr et al., 2020). Stress response-associated programme of uORF translation has been associated with the preferential translation of the immune regulator programmed cell death ligand-1 (PD-L1), and a set of genes related to the development of squamous cell carcinoma (Sendoel et al., 2017).

Active translation of long non-coding RNAs

The concept of actively translated long noncoding RNAs (lncRNAs) is still a controversial and widely debated topic. The main critics of this hypothesis argue that the evidence of lncRNAs translation comes almost exclusively from Ribo-Seq studies. Given that RNAse protected fragments can originate from nonribosomal footprints (Ji et al., 2016), the abundance of footprints observed in certain lncRNAs might be just an artefact. Aebersold et al. (2018b) argue that among all possible amino-acid sequences that could be produced from lncRNAs, only 69 shown proteomic evidence, which, for the majority, was limited to a single peptide match or could be explained by pseudogene missanotation or overlapping exons from adjacent protein coding transcripts. Moreover, most candidate lncRNA-encoded peptides: lack detectable functional protein domain, show lower expression level and lower evolutionary conservation than known protein coding genes (Ji et al., 2015). This, however, provides only a glimpse of a complex system. LncRNAs belongs to a heterogeneous group of 9,640 transcripts (according to ENCODE) with a broad range of biological activity including regulation of transcription, mRNA splicing, sequestration of certain mRNAs or chromatin remodelling (Aebersold et al., 2018b; Statello et al., 2021; Derrien et al., 2012). Annotation of lncRNA has been based predominantly on cDNA alignment to the genome, chromatin signature indicating active transcription and lack of an ORF meeting strict protein-coding criteria (Guttman et al., 2009). Although they do not encode, by convention, canonical proteins, lncRNAs biogenesis is almost indistinguishable from coding mRNAs: they are capped, spliced and polyadenylated (Statello et al., 2021). Pervasive mapping of Ribo-Seq reads to lncRNAs has been reported in several works (Aspden et al., 2014; Bazzini et al., 2014; Ruiz-Orera et al., 2014; Chong et al., 2020; Ouspenskaia et al., 2020; Fields et al., 2015). Indeed, the ribosome occupancy or translation efficiency value is not sufficient to distinguish between true coding and noncoding regions (Guttman et al., 2013; Xiao et al., 2016), but other metrics, such as ORFscore or Ribosome Release Score (RRS), have been developed to assist with this task. Both ORFscore (the strength of frame preference) and RRS (ribosome clearance after stop codon) are highly dependent on the ORF coverage and may underperform for transcripts with low expression values. I expect that, by combining multiple ORF identification algorithms with the strategy of hierarchical merging, it is possible to, firstly, decrease dependency on one metric to score putative ORFs, and secondly, benefit from a large Ribo-Seq dataset to increase sensitivity and specificity of ORF prediction for transcripts with a broad range of expression values. Many lncRNAs-encoded proteins turned out to be functional and have been validated experimentally in humans, fly and mouse (Jackson et al., 2018; Chen et al., 2020). Translated lncRNAs were more hydrophobic with predicted alpha-secondary structure and Kozak sequence around the translation initiation site, but these features varied between studies suggesting strong context or methodology dependency (Ji et al.,

2015; Li and Liu, 2019). Ji et al. (2015) showed that lncRNAs with evident features of active translation were almost exclusively cytoplasmic, had higher conservation scores and evidence of purifying selection of amino-acid sequence.

A consensus model explaining the observations of coding properties of lncRNAs can be proposed. First of all, there is a possibility that in the group of current lncRNAs, there are true protein coding transcripts that have been missed in annotations because of their unusual features. Some *coding* lncRNAs could also form a group of bifunctional RNAs: encoding a protein sequence and biologically important noncoding transcript. A handful of such transcripts has already been observed in human, bacteria, and few model organisms, including *Xenopus*, *Drosophila*, *Zebrafish* (Aebersold et al., 2018b; Hube et al., 2011; Chooniedass-Kothari et al., 2004; Kondo et al., 2010; Ingolia et al., 2011; Kumari and Sampath, 2015).

Dark side of human proteome may shed light on noncanonical translation

The topic of noncanonical translation converges with the concept of *dark proteome*. The dark proteome refers to the portion of the cellular proteome with unknown structure, encoded by *bona-fide* noncoding transcripts or with atypical folding structure (Perdigão et al., 2015). Our preliminary analysis showed that the noncanonical ORFs are enriched for intrinsically disordered regions (IDRs), which is an intriguing observation. IDRs are regions with compositional bias in the amino-acid sequence containing more hydrophilic amino acids and proline residues than structured regions (Dyson and Wright, 2005). IDRs provide a large surface of interaction with frequent short linear motifs (SLIMs), including peptide or nucleic acids binding motifs or sites for post-translational modifications (Tompa et al., 2014; Dinkel et al., 2014). Given their ability to bind and interact with other molecules, IDR-containing proteins, including identified noncanonical ORFs, may act as chaperones or co-factors complementing the function of other proteins (Dyson and Wright, 2005).

CHAPTER 6

PERSPECTIVES

Ribo-Seq as a tool to study genome-wide translation in lymphoid cells

There is no doubt that the technology to study ribosome occupancy with single-nucleotide precision has been a breakthrough for translation studies. Ribo-Seq has been applied to analyse protein synthesis quantitatively, by estimating the translation intensity of a chosen region, or qualitatively, by exploring which portions of the transcriptome undergo active translation.

All Ribo-seq data presented in this thesis has achieved expected quality and reproducibility, similar to in the original protocol (Ingolia et al., 2012). This would not be possible without an efficient bioinformatic workflow. Therefore, RiboStream pipeline, which I have developed for parallel and transparent data processing, was the core stage of this project.

Despite the widespread application of Ribo-Seq in research, a computational workflow has been poorly standardised and the bioinformatic tools developed, specifically for Ribo-Seq, has rarely been used beyond the initial publication. In this thesis, I performed a basic benchmarking of available tools for differential translation analysis to make an informed decision on the strategy for analysing our Ribo-Seq data. The most striking derivative of this analysis was the strong dependency of the performance of the tools on the experimental design. With the most common experimental design in the literature (2 experimental conditions, 2 replicates each) on average only 20% of truly differentially translated genes could be recalled. The number of replicates as high as 8 was just enough to rise the true positive rate to about 70% and stabilise the false positive rate around the desired level of 5%. This suggests that to utilise the full power of Ribo-Seq for differential translation analysis, at least 8 replicates is recommended. Our benchmarking would benefit from an additional set of Ribo-Seq data, that could validate the aforementioned observations. It would be also interesting to dissect characteristics of the transcripts arising as false positives or false negatives, that could assist in the differential translation interpretation.

Translational landscape of GC B-cells malignant transformation

In chapter 3 I have also studied the translational consequences of deregulated expression of the two transcription factors, BCL6 and MYC, in primary GC B-cells. With similar experimental design and biological motivation to find differentially translated genes following the overexpression of an oncogenic transcription factor, I have not observed such profound translational reprogramming as Sendoel et al. (2017). A possible limitation of this study is that Ribo-Seq provides only relative quantification of the ribosome footprint abundance, which may be challenging to interpret when massive changes in gene expression landscape are expected. It would be interesting to perform the same experiment but using polysome fractionation combined with RNA-Seq instead. Up to now, a systematic comparison of Ribo-Seq with polysome fractionation based differential translation analysis has not been performed.

While mRNA levels have directed the majority of changes in gene expression, the differentially translated genes were mainly associated with cellular *housekeeping* functions, such as ribosome biogenesis or oxidative phosphorylation. The adaptive role of translational control in fine tuning of highly energetic metabolic processes and to what extent this may affect the process of malignant transformation is not well understood.

The concept of ribosome heterogeneity is another fascinating topic worth deeper exploration. If BCL6 or MYC-induced changes in the translation intensity of individual ribosomal proteins translates into the stoichiometry of ribosomal proteins incorporated into the ribosome, this could be an essential mechanism of post-transcriptional regulation.

The role of DDX3X in facilitating MYC-driven lymphomagenesis

In chapter 4 I revealed that DDX3X loss-of-function promotes lymphomagenesis by buffering MYC-driven increase in global protein synthesis and proteotoxic stress. I show that DDX3X controls the translation of ribosomal proteins, thus global translation load. Although this involves direct binding of DDX3X to transcripts encoding ribosomal proteins, the exact molecular mechanism is still unclear: whether this is related to RNA unwinding activity of DDX3X, associated with mTOR/LARP1/5'TOP axis or facilitated by one of the other versatile functions of DDX3X.

Another question that remained relating to the role of DDX3X in MYC-driven lymphomas is the mechanism of DDX3Y protein upregulation in transformed B-cells and redundancy of biological activity of DDX3Y and DDX3X in lymphoid cells.

Exploring the role of noncanonical proteome in immunity and tumour immunosurveillance

Finally, in chapter 5 I revealed pervasive translation of noncanonical ORF in lymphoid cells, which is a captivating finding. Although this is not the only study uncovering widespread ribosome occupancy of *bona-fide* noncoding regions, this is one of the largest studies and the first investigating this topic in primary GC B-cells.

An introductory analysis presented in this thesis undoubtedly has a hypothesis-generating flavour. The most burning questions related to noncanonical translation involve the extension of the protein level expression, the interplay between uORF and main ORF translation, and the frequency of potential functional protein domains and short linear motifs, which could shed light on the biological role of synthesised micropeptides.

The analysis of three external immunopeptidomics data revealed that between 2 and 5% of MHC-bound peptides (MAPs) could originate from predicted noncanonical ORFs. To what extent this refers to the immunogenic function of noncanonical translation or reflects the inferior performance of other proteomic methods, remains to be addressed.

Bibliography

Abate, F., Ambrosio, M. R., Mundo, L., Laginestra, M. A., Fuligni, F., Rossi, M., Zairis, S., Gazaneo, S., De Falco, G., Lazzi, S., et al. (2015). Distinct viral and mutational spectrum of endemic Burkitt lymphoma. *PLoS pathogens*, 11(10):e1005158.

Aebersold, R., Agar, J. N., Amster, I. J., Baker, M. S., Bertozzi, C. R., Boja, E. S., Costello, C. E., Cravatt, B. F., Fenselau, C., Garcia, B. A., et al. (2018a). How many human proteoforms are there? *Nature Chemical Biology*, 14(3):206.

Aebersold, R., Agar, J. N., Amster, I. J., Baker, M. S., Bertozzi, C. R., Boja, E. S., Costello, C. E., Cravatt, B. F., Fenselau, C., Garcia, B. A., et al. (2018b). How many human proteoforms are there? *Nature Chemical Biology*, 14(3):206–214.

Aitken, C. E. and Lorsch, J. R. (2012). A mechanistic overview of translation initiation in eukaryotes. *Nature Structural and Molecular Biology*, 19(6):568–576.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511.

Alkallas, R., Lajoie, M., Moldoveanu, D., Hoang, K. V., Lefrançois, P., Lingrand, M., Ahanfeshar-Adams, M., Watters, K., Spatz, A., Zippin, J. H., et al. (2020). Multi-omic analysis reveals significantly mutated genes and DDX3X as a sex-specific tumor suppressor in cutaneous melanoma. *Nature Cancer*, 1(6):635–652.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Andrews, S. et al. (2017). FastQC: a quality control tool for high throughput sequence data. 2010.

Ansell, S. M., Lesokhin, A. M., Borrello, I., Halwani, A., Scott, E. C., Gutierrez, M., Schuster, S. J., Millenson, M. M., Cattry, D., Freeman, G. J., et al. (2015). PD-1 blockade with nivolumab in relapsed or refractory Hodgkin's lymphoma. *New England Journal of Medicine*, 372(4):311–319.

Aspden, J. L., Eyre-Walker, Y. C., Phillips, R. J., Amin, U., Mumtaz, M. A. S., Brocard, M., and Couso, J.-P. (2014). Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *eLife*, 3:e03528.

Aviner, R. (2020). The science of puromycin: From studies of ribosome function to applications in biotechnology. *Computational and Structural Biotechnology Journal*, 18:1074–1083.

Aviner, R., Geiger, T., and Elroy-Stein, O. (2013). Novel proteomic approach (PUNCH-P) reveals cell cycle-specific fluctuations in mRNA translation. *Genes & development*, 27(16):1834–1844.

Babaian, A., Rothe, K., Girodat, D., Minia, I., Djondovic, S., Milek, M., Miko, S. E. S., Wieden, H.-J., Landthaler, M., Morin, G. B., et al. (2020). Loss of m1acp3ψ ribosomal RNA modification is a major feature of cancer. *Cell Reports*, 31(5):107611.

Barna, M., Pusic, A., Zollo, O., Costa, M., Kondrashov, N., Rego, E., Rao, P. H., and Ruggero, D. (2008). Suppression of Myc oncogenic activity by ribosomal protein haploinsufficiency. *Nature*, 456(7224):971–975.

Basso, K. and Dalla-Favera, R. (2015). Germinal centres and B cell lymphomagenesis. *Nature Reviews Immunology*, 15(3):172–184.

Bastide, A. and David, A. (2018). The ribosome, (slow) beating heart of cancer (stem) cell. *Oncogenesis*, 7(4):1–13.

Battle, A., Khan, Z., Wang, S. H., Mitrano, A., Ford, M. J., Pritchard, J. K., and Gilad, Y. (2015). Impact of regulatory variation from RNA to protein. *Science*, 347(6222):664–667.

Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., Vejnar, C. E., Lee, M. T., Rajewsky, N., Walther, T. C., et al. (2014). Identification of small ORF s in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO journal*, 33(9):981–993.

Beadle, G. W. and Tatum, E. L. (1941). Genetic control of biochemical reactions in Neurospora. *Proceedings of the National Academy of Sciences of the United States of America*, 27(11):499.

Bekaert, M., Ivanov, I. P., Atkins, J. F., and Baranov, P. V. (2008). Ornithine decarboxylase antizyme finder (OAF): fast and reliable detection of antizymes with frameshifts in mRNAs. *BMC Bioinformatics*, 9(1):1–10.

Bekker-Jensen, D. B., Kelstrup, C. D., Batth, T. S., Larsen, S. C., Haldrup, C., Bramsen, J. B., Sørensen, K. D., Høyer, S., Ørntoft, T. F., Andersen, C. L., et al. (2017). An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Systems*, 4(6):587–599.

Berletch, J. B., Yang, F., Xu, J., Carrel, L., and Disteche, C. M. (2011). Genes that escape from X inactivation. *Human Genetics*, 130(2):237–245.

Bhat, M., Robichaud, N., Hulea, L., Sonenberg, N., Pelletier, J., and Topisirovic, I. (2015). Targeting the translation machinery in cancer. *Nature Reviews Drug discovery*, 14(4):261–278.

Blakeley, P., Overton, I. M., and Hubbard, S. J. (2012). Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *Journal of Proteome Research*, 11(11):5221–5234.

Bol, G. M., Vesuna, F., Xie, M., Zeng, J., Aziz, K., Gandhi, N., Levine, A., Irving, A., Korz, D., Tantravedi, S., et al. (2015). Targeting DDX3 with a small molecule inhibitor for lung cancer therapy. *EMBO Molecular Medicine*, 7(5):648–669.

Boon, K., Caron, H. N., Van Asperen, R., Valentijn, L., Hermus, M.-C., Van Sluis, P., Roobeek, I., Weis, I., Voute, P., Schwab, M., et al. (2001). N-myc enhances the expression of a large set of genes functioning in ribosome biogenesis and protein synthesis. *The EMBO Journal*, 20(6):1383–1393.

Bourgeois, C. F., Mortreux, F., and Auboeuf, D. (2016). The multiple functions of RNA helicases as drivers and regulators of gene expression. *Nature Reviews Molecular Cell Biology*, 17(7):426–438.

Bouska, A., Bi, C., Lone, W., Zhang, W., Kedwaii, A., Heavican, T., Lachel, C. M., Yu, J., Ferro, R., Eldorghamy, N., et al. (2017). Adult high-grade b-cell lymphoma with Burkitt lymphoma signature: genomic features and potential therapeutic targets. *Blood*, 130(16):1819–1831.

Brai, A., Fazi, R., Tintori, C., Zamperini, C., Bugli, F., Sanguinetti, M., Stigliano, E., Esté, J., Badia, R., Franco, S., et al. (2016). Human DDX3 protein is a valuable target to develop broad spectrum antiviral agents. *Proceedings of the National Academy of Sciences*, 113(19):5388–5393.

Brent, M. R. (2005). Genome annotation past, present, and future: how to define an ORF at each locus. *Genome research*, 15(12):1777–1786.

Brunet, M. A., Lucier, J.-F., Levesque, M., Leblanc, S., Jacques, J.-F., Al-Saedi, H. R., Guilloy, N., Grenier, F., Avino, M., Fournier, I., et al. (2021). OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Research*, 49(D1):D380–D388.

Buchan, J. R. and Parker, R. (2009). Eukaryotic stress granules: the ins and outs of translation. *Molecular Cell*, 36(6):932–941.

Buttgereit, F. and Brand, M. D. (1995). A hierarchy of ATP-consuming processes in mammalian cells. *Biochemical Journal*, 312(1):163–167.

Caeser, R., Di Re, M., Krupka, J. A., Gao, J., Lara-Chica, M., Dias, J. M., Cooke, S. L., Fenner, R., Usheva, Z., Runge, H. F., et al. (2019). Genetic modification of primary human B cells to model high-grade lymphoma. *Nature Communications*, 10(1):1–16.

Cai, X., Gao, L., Teng, L., Ge, J., Oo, Z. M., Kumar, A. R., Gilliland, D. G., Mason, P. J., Tan, K., and Speck, N. A. (2015). Runx1 deficiency decreases ribosome biogenesis and confers stress resistance to hematopoietic stem and progenitor cells. *Cell Stem Cell*, 17(2):165–177.

Calado, D. P., Sasaki, Y., Godinho, S. A., Pellerin, A., Köchert, K., Sleckman, B. P., De Alborán, I. M., Janz, M., Rodig, S., and Rajewsky, K. (2012). The cell-cycle regulator c-Myc is essential for the formation and maintenance of germinal centers. *Nature Immunology*, 13(11):1092–1100.

Calviello, L. and Ohler, U. (2017). Beyond read-counts: Ribo-seq data analysis to understand the functions of the transcriptome. *Trends in Genetics*, 33(10):728–744.

Calviello, L., Venkataramanan, S., Rogowski, K. J., Wyler, E., Wilkins, K., Tejura, M., Thai, B., Krol, J., Filipowicz, W., Landthaler, M., et al. (2021). DDX3 depletion represses translation of mRNAs with complex 5 UTRs. *Nucleic Acids Research*, 49(9):5336–5350.

Calvo, S. E., Pagliarini, D. J., and Mootha, V. K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proceedings of the National Academy of Sciences*, 106(18):7507–7512.

Cannizzaro, E., Bannister, A. J., Han, N., Alendar, A., and Kouzarides, T. (2018). DDX3X RNA helicase affects breast cancer cell cycle progression by regulating expression of KLF4. *FEBS Letters*, 592(13):2308–2322.

Casey, S. C., Baylot, V., and Felsher, D. W. (2018). The MYC oncogene is a global regulator of the immune response. *Blood, The Journal of the American Society of Hematology*, 131(18):2007–2015.

Casey, S. C., Tong, L., Li, Y., Do, R., Walz, S., Fitzgerald, K. N., Gouw, A. M., Baylot, V., Gütgemann, I., Eilers, M., et al. (2016). MYC regulates the antitumor immune response through CD47 and PD-L1. *Science*, 352(6282):227–231.

Cech, T. R. (2000). The ribosome is a ribozyme. *Science*, 289(5481):878–879.

Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., et al. (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, 30(10):918–920.

Chapuy, B., Stewart, C., Dunford, A. J., Kim, J., Kamburov, A., Redd, R. A., Lawrence, M. S., Roemer, M. G., Li, A. J., Ziepert, M., et al. (2018). Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nature Medicine*, 24(5):679–690.

Chen, H., Liu, H., and Qing, G. (2018). Targeting oncogenic Myc as a strategy for cancer treatment. *Signal transduction and targeted therapy*, 3(1):1–7.

Chen, J., Brunner, A.-D., Cogan, J. Z., Nuñez, J. K., Fields, A. P., Adamson, B., Itzhak, D. N., Li, J. Y., Mann, M., Leonetti, M. D., et al. (2020). Pervasive functional translation of noncanonical human open reading frames. *Science*, 367(6482):1140–1146.

Chew, G.-L., Pauli, A., and Schier, A. F. (2016). Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nature Communications*, 7(1):1–10.

Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., Rad, R., Huttlin, E. L., and Gygi, S. P. (2015). A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature Biotechnology*, 33(7):743–749.

Chong, C., Müller, M., Pak, H., Harnett, D., Huber, F., Grun, D., Leleu, M., Auger, A., Arnaud, M., Stevenson, B. J., et al. (2020). Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nature Communications*, 11(1):1–21.

Chooniedass-Kothari, S., Emberley, E., Hamedani, M., Troup, S., Wang, X., Czosnek, A., Hube, F., Mutawe, M., Watson, P., and Leygue, E. (2004). The steroid receptor RNA activator is the first functional RNA encoding a protein. *FEBS Letters*, 566(1-3):43–47.

Chothani, S., Adami, E., Ouyang, J. F., Viswanathan, S., Hubner, N., Cook, S. A., Schafer, S., and Rackham, O. J. (2019). deltaTE: detection of translationally regulated genes by integrative analysis of Ribo-seq and RNA-seq data. *Current Protocols in Molecular Biology*, 129(1):e108.

Churbanov, A., Rogozin, I. B., Babenko, V. N., Ali, H., and Koonin, E. V. (2005). Evolutionary conservation suggests a regulatory function of AUG triplets in 5-UTRs of eukaryotic genes. *Nucleic Acids Research*, 33(17):5512–5520.

Ci, W., Polo, J. M., Cerchietti, L., Shaknovich, R., Wang, L., Yang, S. N., Ye, K., Farinha, P., Horsman, D. E., Gascoyne, R. D., et al. (2009). The BCL6 transcriptional program features repression of multiple oncogenes in primary B cells and is deregulated in DLBCL. *Blood*, 113(22):5536–5548.

Clarke, H. J., Chambers, J. E., Liniker, E., and Marciniak, S. J. (2014). Endoplasmic reticulum stress in malignancy. *Cancer Cell*, 25(5):563–573.

Consortium, G. et al. (2020). The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330.

Costa, L. J., Xavier, A. C., Wahlquist, A. E., and Hill, E. G. (2013). Trends in survival of patients with Burkitt lymphoma/leukemia in the USA: an analysis of 3691 cases. *Blood*, 121(24):4861–4866.

Costa-Mattioli, M. and Walter, P. (2020). The integrated stress response: From mechanism to disease. *Science*, 368(6489).

Cotton, A. M., Price, E. M., Jones, M. J., Balaton, B. P., Kobor, M. S., and Brown, C. J. (2015). Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and x-chromosome inactivation. *Human Molecular Genetics*, 24(6):1528–1539.

Cox, J. and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372.

Crick, F. H. (1958). On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8.

Cruciat, C.-M., Dolde, C., De Groot, R. E., Ohkawara, B., Reinhard, C., Korswagen, H. C., and Niehrs, C. (2013). RNA helicase DDX3 is a regulatory subunit of casein kinase 1 in Wnt–β-catenin signaling. *Science*, 339(6126):1436–1441.

Cucco, F., Barrans, S., Sha, C., Clipson, A., Crouch, S., Dobson, R., Chen, Z., Thompson, J. S., Care, M. A., Cummin, T., et al. (2020). Distinct genetic changes reveal evolutionary history and heterogeneous molecular grade of DLBCL with MYC/BCL2 double-hit. *Leukemia*, 34(5):1329–1341.

Cuevas, M. V. R., Hardy, M.-P., Hollỳ, J., Bonneil, É., Durette, C., Courcelles, M., Lanoix, J., Côté, C., Staudt, L. M., Lemieux, S., et al. (2021). Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Reports*, 34(10):108815.

Culjkovic-Kraljacic, B., Fernando, T. M., Marullo, R., Calvo-Vidal, N., Verma, A., Yang, S., Tabbò, F., Gaudiano, M., Zahreddine, H., Goldstein, R. L., et al. (2016). Combinatorial targeting of nuclear export and translation of RNA inhibits aggressive b-cell lymphomas. *Blood*, 127(7):858–868.

Cutmore, Krupka, Hodson (2022). Molecular profiling in diffuse large b cell lymphoma – challenges and opportunities. *Under review.*

Dai, M.-S., Sun, X.-X., and Lu, H. (2010). Ribosomal protein L11 associates with c-Myc at 5S rRNA and tRNA genes and regulates their expression. *Journal of Biological Chemistry*, 285(17):12587–12594.

Dang, C. V. (2012). Myc on the path to cancer. *Cell*, 149(1):22–35.

de Loubresse, N. G., Prokhorova, I., Holtkamp, W., Rodnina, M. V., Yusupova, G., and Yusupov, M. (2014). Structural basis for the inhibition of the eukaryotic ribosome. *Nature*, 513(7519):517–522.

De Silva, N. S. and Klein, U. (2015). Dynamics of B cells in germinal centres. *Nature Reviews Immunology*, 15(3):137–148.

Deeb, S. J., Cox, J., Schmidt-Supprian, M., and Mann, M. (2014). N-linked glycosylation enrichment for in-depth cell surface proteomics of diffuse large B-cell lymphoma subtypes. *Molecular & Cellular Proteomics*, 13(1):240–251.

Deeb, S. J., D'Souza, R. C., Cox, J., Schmidt-Supprian, M., and Mann, M. (2012). Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles. *Molecular & Cellular Proteomics*, 11(5):77–89.

Delaidelli, A., Leprivier, G., and Sorensen, P. H. (2017). eEF2K protects MYCN-amplified cells from starvation. *Cell Cycle*, 16(18):1633.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., et al. (2012). The GENCODE v7 catalog of human long noncoding rnas: analysis of their gene structure, evolution, and expression. *Genome Research*, 22(9):1775–1789.

Dersh, D., Hollỳ, J., and Yewdell, J. W. (2021). A few good peptides: MHC class I-based cancer immunosurveillance and immunoevasion. *Nature Reviews Immunology*, 21(2):116–128.

Desnoyers, G., Frost, L. D., Courteau, L., Wall, M. L., and Lewis, S. M. (2015). Decreased eIF3e expression can mediate epithelial-to-mesenchymal transition through activation of the TGFβ signaling pathway. *Molecular Cancer Research*, 13(10):1421–1430.

Deutsch, E. W., Csordas, A., Sun, Z., Jarnuczak, A., Perez-Riverol, Y., Ternent, T., Campbell, D. S., Bernal-Llinares, M., Okuda, S., Kawano, S., et al. (2016). The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Research*, page gkw936.

Dinkel, H., Van Roey, K., Michael, S., Davey, N. E., Weatheritt, R. J., Born, D., Speck, T., Krüger, D., Grebnev, G., Kubań, M., et al. (2014). The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Research*, 42(D1):D259–D266.

Dittmar, K. A., Goodenbour, J. M., and Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. *PLoS genetics*, 2(12):e221.

Ditton, H., Zimmer, J., Kamp, C., Rajpert-De Meyts, E., and Vogt, P. (2004). The AZFa gene dby (DDX3Y) is widely transcribed but the protein is limited to the male germ cells by translation control. *Human Molecular Genetics*, 13(19):2333–2341.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, 34(2):184–191.

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J., and Root, D. E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. *Nature Biotechnology*, 32(12):1262–1267.

Dominguez-Sola, D., Victora, G. D., Ying, C. Y., Phan, R. T., Saito, M., Nussenzweig, M. C., and Dalla-Favera, R. (2012). The proto-oncogene MYC is required for selection in the germinal center and cyclic reentry. *Nature Immunology*, 13(11):1083–1091.

Dresios, J., Chappell, S. A., Zhou, W., and Mauro, V. P. (2006). An mRNA-rRNA base-pairing mechanism for translation initiation in eukaryotes. *Nature structural & Molecular Biology*, 13(1):30–34.

Duncan, C. D. and Mata, J. (2017). Effects of cycloheximide on the interpretation of ribosome profiling experiments in Schizosaccharomyces pombe. *Scientific Reports*, 7(1):1–11.

Dunford, A., Weinstock, D. M., Savova, V., Schumacher, S. E., Cleary, J. P., Yoda, A., Sullivan, T. J., Hess, J. M., Gimelbrant, A. A., Beroukhim, R., et al. (2017). Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias. *Nature Genetics*, 49(1):10–16.

Dutt, S., Narla, A., Lin, K., Mullally, A., Abayasekara, N., Megerdichian, C., Wilson, F. H., Currie, T., Khanna-Gupta, A., Berliner, N., et al. (2011). Haploinsufficiency for ribosomal protein genes causes selective activation of p53 in human erythroid progenitor cells. *Blood*, 117(9):2567–2576.

Dyson, H. J. and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell biology*, 6(3):197–208.

Eng, J. K., Jahan, T. A., and Hoopmann, M. R. (2013). Comet: an open-source MS/MS sequence database search tool. *Proteomics*, 13(1):22–24.

Ennishi, D., Jiang, A., Boyle, M., Collinge, B., Grande, B. M., Ben-Neriah, S., Rushton, C., Tang, J., Thomas, N., Slack, G. W., et al. (2019). Double-hit gene expression signature defines a distinct subgroup of germinal center b-cell-like diffuse large b-cell lymphoma. *Journal of Clinical Oncology*, 37(3):190.

Eswarappa, S. M., Potdar, A. A., Koch, W. J., Fan, Y., Vasu, K., Lindner, D., Willard, B., Graham, L. M., DiCorleto, P. E., and Fox, P. L. (2014). Programmed translational readthrough generates antiangiogenic VEGF-Ax. *Cell*, 157(7):1605–1618.

Etzioni, A. and Ochs, H. D. (2004). The hyper IgM syndrome—an evolving story. *Pediatric research*, 56(4):519–525.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048.

Ferlin, A., Moro, E., Rossi, A., Dallapiccola, B., and Foresta, C. (2003). The human Y chromosome's azoospermia factor b (azfb) region: sequence, structure, and deletion analysis in infertile men. *Journal of Medical Genetics*, 40(1):18–24.

Fields, A. P., Rodriguez, E. H., Jovanovic, M., Stern-Ginossar, N., Haas, B. J., Mertins, P., Raychowdhury, R., Hacohen, N., Carr, S. A., Ingolia, N. T., et al. (2015). A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Molecular Cell*, 60(5):816–827.

Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature Reviews Genetics*, 9(2):102–114.

Floor, S. N., Condon, K. J., Sharma, D., Jankowsky, E., and Doudna, J. A. (2016). Autoinhibitory interdomain interactions and subfamily-specific extensions redefine the catalytic core of the human dead-box protein ddx3. *Journal of Biological Chemistry*, 291(5):2412–2421.

Foresta, C., Ferlin, A., and Moro, E. (2000a). Deletion and expression analysis of AZFa genes on the human Y chromosome revealed a major role for DBY in male infertility. *Human Molecular Genetics*, 9(8):1161–1169.

Foresta, C., Ferlin, A., and Moro, E. (2000b). Deletion and expression analysis of AZFa genes on the human Y chromosome revealed a major role for DBY in male infertility. *Human Molecular Genetics*, 9(8):1161–1169.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773.

Furic, L., Rong, L., Larsson, O., Koumakpayi, I. H., Yoshida, K., Brueschke, A., Petroulakis, E., Robichaud, N., Pollak, M., Gaboury, L. A., et al. (2010). eIF4E phosphorylation promotes tumorigenesis and is associated with prostate cancer progression. *Proceedings of the National Academy of Sciences*, 107(32):14134–14139.

Gani, R. (1976). The nucleoli of cultured human lymphocytes: I. nucleolar morphology in relation to transformation and the DNA cycle. *Experimental Cell Research*, 97(2):249–258.

Gao, H., Sun, X., and Rao, Y. (2020). PROTAC technology: opportunities and challenges. *ACS Medicinal Chemistry Letters*, 11(3):237–240.

Gao, X., Wan, J., Liu, B., Ma, M., Shen, B., and Qian, S.-B. (2015). Quantitative profiling of initiating ribosomes in vivo. *Nature Methods*, 12(2):147–153.

Gao, Z., Herrera-Carrillo, E., and Berkhout, B. (2018). Delineation of the exact transcription termination signal for type 3 polymerase III. *Molecular Therapy-Nucleic Acids*, 10:36–44.

Geissler, R., Golbik, R. P., and Behrens, S.-E. (2012). The DEAD-box helicase DDX3 supports the assembly of functional 80s ribosomes. *Nucleic Acids Research*, 40(11):4998–5011.

Genuth, N. R. and Barna, M. (2018). The discovery of ribosome heterogeneity and its implications for gene regulation and organisms life. *Molecular Cell*, 71(3):364–374.

Gerashchenko, M. V. and Gladyshev, V. N. (2014). Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Research*, 42(17):e134–e134.

Gerashchenko, M. V. and Gladyshev, V. N. (2017). Ribonuclease selection for ribosome profiling. *Nucleic Acids Research*, 45(2):e6–e6.

Ghaddar, N., Wang, S., Woodvine, B., Krishnamoorthy, J., van Hoef, V., Darini, C., Kazimierczak, U., Ah-Son, N., Popper, H., Johnson, M., et al. (2021). The integrated stress response is tumorigenic and constitutes a therapeutic liability in KRAS-driven lung cancer. *Nature Communications*, 12(1):1–15.

Gingras, A.-C., Raught, B., and Sonenberg, N. (1999). eIF4 initiation factors: effectors of mRNA recruitment to ribosomes and regulators of translation. *Annual review of biochemistry*, 68(1):913–963.

God, J. M., Cameron, C., Figueroa, J., Amria, S., Hossain, A., Kempkes, B., Bornkamm, G. W., Stuart, R. K., Blum, J. S., and Haque, A. (2015). Elevation of c-MYC disrupts HLA class II–mediated immune recognition of human B cell tumors. *The Journal of Immunology*, 194(4):1434–1445.

Gong, Krupka, Gao, J., Grigoropoulos, N. F., Screen, M., Usheva, Z., Cucco, F., Barrans, S., Painter, D., Mohammed, M., et al. (2021). Sequential inverse dysregulation of the RNA helicases DDX3X and DDX3Y facilitates MYC-driven lymphomagenesis.

Good-Jacobson, K. L., Szumilas, C. G., Chen, L., Sharpe, A. H., Tomayko, M. M., and Shlomchik, M. J. (2010). PD-1 regulates germinal center B cell survival and the formation and affinity of long-lived plasma cells. *Nature Immunology*, 11(6):535–542.

Goodman, A., Patel, S. P., and Kurzrock, R. (2017). PD-1–PD-L1 immune-checkpoint blockade in B-cell lymphomas. *Nature Reviews Clinical Oncology*, 14(4):203–220.

Grande, B. M., Gerhard, D. S., Jiang, A., Griner, N. B., Abramson, J. S., Alexander, T. B., Allen, H., Ayers, L. W., Bethony, J. M., Bhatia, K., et al. (2019). Genome-wide discovery of somatic coding and noncoding mutations in pediatric endemic and sporadic Burkitt lymphoma. *Blood*, 133(12):1313–1324.

Grandori, C., Gomez-Roman, N., Felton-Edkins, Z. A., Ngouenet, C., Galloway, D. A., Eisenman, R. N., and White, R. J. (2005). c-Myc binds to human ribosomal DNA and stimulates transcription of rRNA genes by RNA polymerase I. *Nature Cell Biology*, 7(3):311–318.

Green, M. R., Monti, S., Rodig, S. J., Juszczynski, P., Currie, T., O'Donnell, E., Chapuy, B., Takeyama, K., Neuberg, D., Golub, T. R., et al. (2010). Integrative analysis reveals selective 9p24. 1 amplification, increased PD-1 ligand expression, and further induction via JAK2 in nodular sclerosing Hodgkin lymphoma and primary mediastinal large b-cell lymphoma. *Blood*, 116(17):3268–3277.

Griss, J., Perez-Riverol, Y., Lewis, S., Tabb, D. L., Dianes, J. A., Del-Toro, N., Rurik, M., Walzer, M., Kohlbacher, O., Hermjakob, H., et al. (2016). Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nature Methods*, 13(8):651–656.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458(7235):223–227.

Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., and Lander, E. S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell*, 154(1):240–251.

Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., and Zavolan, M. (2021). CLIP and complementary methods. *Nature Reviews Methods Primers*, 1(1):1–23.

Hämmerl, L., Colombet, M., Rochford, R., Ogwang, D. M., and Parkin, D. M. (2019). The burden of Burkitt lymphoma in Africa. *Infectious Agents and Cancer*, 14(1):1–6.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, 144(5):646–674.

Hanson, G. and Coller, J. (2018). Codon optimality, bias and usage in translation and mRNA decay. *Nature reviews Molecular cell biology*, 19(1):20.

Hariri, F., Arguello, M., Volpon, L., Culjkovic-Kraljacic, B., Nielsen, T. H., Hiscott, J., Mann, K. K., and Borden, K. L. (2013). The eukaryotic translation initiation factor eif4e is a direct transcriptional target of NF-κB and is aberrantly regulated in acute myeloid leukemia. *Leukemia*, 27(10):2047–2055.

He, Y., Zhang, D., Yang, Y., Wang, X., Zhao, X., Zhang, P., Zhu, H., Xu, N., and Liang, S. (2018). A double-edged function of DDX3, as an oncogene or tumor suppressor, in cancer progression. *Oncology reports*, 39(3):883–892.

Hellen, C. U. (2018). Translation termination and ribosome recycling in eukaryotes. *Cold Spring Harbor perspectives in biology*, 10(10):a032656.

Helmrich, A., Ballarino, M., and Tora, L. (2011). Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Molecular cell*, 44(6):966–977.

Henderson, A., Warburton, D., and Atwood, K. (1972). Location of ribosomal DNA in the human chromosome complement. *Proceedings of the National Academy of Sciences*, 69(11):3394–3398.

Hershey, J. W. B., Sonenberg, N., and Mathews, M. B. (2012). Principles of Translational Control: An Overview. *Cold Spring Harbor Perspectives in Biology*, 4(12):a011528–a011528.

Hetz, C. (2012). The unfolded protein response: controlling cell fate decisions under ER stress and beyond. *Nature reviews Molecular cell biology*, 13(2):89–102.

Hetz, C., Chevet, E., and Oakes, S. A. (2015). Proteostasis control by the unfolded protein response. *Nature Cell Biology*, 17(7):829–838.

Hilliker, A., Gao, Z., Jankowsky, E., and Parker, R. (2011). The DEAD-box protein Ded1 modulates translation by the formation and resolution of an eIF4F-mRNA complex. *Molecular Cell*, 43(6):962–972.

Hinnebusch, A. G. (2005). Translational regulation of GCN4 and the general amino acid control of yeast. *Annu. Rev. Microbiol.*, 59:407–450.

Hinnebusch, A. G. (2014). The scanning mechanism of eukaryotic translation initiation. *Annual Review of Biochemistry*, 83:779–812.

Hinnebusch, A. G., Ivanov, I. P., and Sonenberg, N. (2016). Translational control by 5-untranslated regions of eukaryotic mRNAs. *Science*, 352(6292):1413–1416.

Ho, J. S., Ma, W., Mao, D. Y., and Benchimol, S. (2005). p53-dependent transcriptional repression of c-myc is required for G1 cell cycle arrest. *Molecular and Cellular Biology*, 25(17):7423–7431.

Ho, Y., Li, X., Jamison, S., Harding, H. P., McKinnon, P. J., Ron, D., and Lin, W. (2016). PERK activation promotes medulloblastoma tumorigenesis by attenuating premalignant granule cell precursor apoptosis. *The American Journal of Pathology*, 186(7):1939–1951.

Horvilleur, E., Sbarrato, T., Hill, K., Spriggs, R., Screen, M., Goodrem, P., Sawicka, K., Chaplin, L., Touriol, C., Packham, G., et al. (2014). A role for eukaryotic initiation factor 4B overexpression in the pathogenesis of diffuse large b-cell lymphoma. *Leukemia*, 28(5):1092–1102.

Horvilleur, E., Wilson, L. A., and Willis, A. E. (2010). Translation deregulation in B-cell lymphomas.

Howden, A. J., Geoghegan, V., Katsch, K., Efstathiou, G., Bhushan, B., Boutureira, O., Thomas, B., Trudgian, D. C., Kessler, B. M., Dieterich, D. C., et al. (2013). QuaNCAT: quantitating proteome dynamics in primary cells. *Nature Methods*, 10(4):343–346.

Hsieh, A. C., Liu, Y., Edlind, M. P., Ingolia, N. T., Janes, M. R., Sher, A., Shi, E. Y., Stumpf, C. R., Christensen, C., Bonham, M. J., Wang, S., Ren, P., Martin, M., Jessen, K., Feldman, M. E., Weissman, J. S., Shokat, K. M., Rommel, C., and Ruggero, D. (2012). The translational landscape of mTOR signalling steers cancer initiation and metastasis. *Nature*, 485(7396):55–61.

Hu, F., Lu, J., Munoz, M. D., Saveliev, A., and Turner, M. (2021). ORFLine: a bioinformatic pipeline to prioritise small open reading frames identifies candidate secreted small proteins from lymphocytes. *bioRxiv*.

Huang, L., Luo, R., Li, J., Wang, D., Zhang, Y., Liu, L., Zhang, N., Xu, X., Lu, B., and Zhao, K. (2020). β-catenin promotes NLRP3 inflammasome activation via increasing the association between NLRP3 and ASC. *Molecular Immunology*, 121:186–194.

Hube, F., Velasco, G., Rollin, J., Furling, D., and Francastel, C. (2011). Steroid receptor RNA activator protein binds to and counteracts SRA RNA-mediated activation of MyoD and muscle differentiation. *Nucleic Acids Research*, 39(2):513–525.

Hussmann, J. A., Patchett, S., Johnson, A., Sawyer, S., and Press, W. H. (2015). Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. *PLoS Genetics*, 11(12):e1005732.

Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M., and Weissman, J. S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature Protocols*, 7(8):1534–1550.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223.

Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802.

Iossifov, I., O'roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, 515(7526):216–221.

Ivanov, P., Emara, M. M., Villen, J., Gygi, S. P., and Anderson, P. (2011). Angiogenin-Induced tRNA Fragments Inhibit Translation Initiation. *Molecular Cell*, 43(4):613–623.

Ivanov, P., O'Day, E., Emara, M. M., Wagner, G., Lieberman, J., and Anderson, P. (2014). G-quadruplex structures contribute to the neuroprotective effects of angiogenin-induced trna fragments. *Proceedings of the National Academy of Sciences*, 111(51):18201–18206.

Iwasaki, S. and Ingolia, N. T. (2017). The growing toolbox for protein synthesis studies. *Trends in biochemical sciences*, 42(8):612–624.

Jackson, R., Kroehling, L., Khitun, A., Bailis, W., Jarret, A., York, A. G., Khan, O. M., Brewer, J. R., Skadow, M. H., Duizer, C., et al. (2018). The translation of non-canonical open reading frames controls mucosal immunity. *Nature*, 564(7736):434–438.

Jackson, R. J., Hellen, C. U., and Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews Molecular cell biology*, 11(2):113–127.

Jain, S., Wheeler, J. R., Walters, R. W., Agrawal, A., Barsic, A., and Parker, R. (2016). ATPase-modulated stress granules contain a diverse proteome and substructure. *Cell*, 164(3):487–498.

Jaroszynski, L., Zimmer, J., Fietz, D., Bergmann, M., Kliesch, S., and Vogt, P. (2011). Translational control of the AZFa gene DDX3Y by 5 UTR exon-T extension. *International journal of andrology*, 34(4pt1):313–326.

Ji, Z. (2018). RibORF: identifying genome-wide translated open reading frames using ribosome profiling. *Current Protocols in Molecular Biology*, 124(1):e67.

Ji, Z., Song, R., Huang, H., Regev, A., and Struhl, K. (2016). Transcriptome-scale rnase-footprinting of rna-protein complexes. *Nature Biotechnology*, 34(4):410–413.

Ji, Z., Song, R., Regev, A., and Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *elife*, 4:e08890.

Jiang, L., Gu, Z.-H., Yan, Z.-X., Zhao, X., Xie, Y.-Y., Zhang, Z.-G., Pan, C.-M., Hu, Y., Cai, C.-P., Dong, Y., et al. (2015). Exome sequencing identifies somatic mutations of DDX3X in natural killer/T-cell lymphoma. *Nature Genetics*, 47(9):1061–1066.

Johannes, G., Carter, M. S., Eisen, M. B., Brown, P. O., and Sarnow, P. (1999). Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. *Proceedings of the National Academy of Sciences*, 96(23):13118–13123.

Johnson, L. F., Levis, R., Abelson, H. T., Green, H., and Penman, S. (1976). Changes in RNA in relation to growth of the fibroblast. iv. alterations in the production and processing of mRNA and rrna in resting and growing cells. *The Journal of Cell Biology*, 71(3):933–938.

Johnson-Kerner, B., Blok, L. S., Suit, L., Thomas, J., Kleefstra, T., and Sherr, E. H. (2020). DDX3X-related neurodevelopmental disorder. *GeneReviews®[Internet]*.

Johnston, H. E., Carter, M. J., Larrayoz, M., Clarke, J., Garbis, S. D., Oscier, D., Strefford, J. C., Steele, A. J., Walewska, R., and Cragg, M. S. (2018). Proteomics profiling of CLL versus healthy B-cells identifies putative therapeutic targets and a subtype-independent signature of spliceosome dysregulation. *Molecular & Cellular Proteomics*, 17(4):776–791.

Johnstone, T. G., Bazzini, A. A., and Giraldez, A. J. (2016). Upstream ORF s are prevalent translational repressors in vertebrates. *The EMBO journal*, 35(7):706–723.

Jones, D. T., Jäger, N., Kool, M., Zichner, T., Hutter, B., Sultan, M., Cho, Y.-J., Pugh, T. J., Hovestadt, V., Stütz, A. M., et al. (2012). Dissecting the genomic complexity underlying medulloblastoma. *Nature*, 488(7409):100–105.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., et al. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(suppl_1):D428–D432.

Jovanovic, M., Rooney, M. S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E. H., Fields, A. P., Schwartz, S., Raychowdhury, R., et al. (2015). Dynamic profiling of the protein life cycle in response to pathogens. *Science*, 347(6226).

Kampen, K. R., Sulima, S. O., Vereecke, S., and De Keersmaecker, K. (2020). Hallmarks of ribosomopathies. *Nucleic acids research*, 48(3):1013–1028.

Kapadia, B., Nanaji, N. M., Bhalla, K., Bhandary, B., Lapidus, R., Beheshti, A., Evens, A. M., and Gartenhaus, R. B. (2018). Fatty acid synthase induced S6Kinase facilitates USP11-eIF4B complex formation for sustained oncogenic translation in DLBCL. *Nature Communications*, 9(1):1–15.

Karginov, F. V. and Hannon, G. J. (2013). Remodeling of Ago2–mRNA interactions upon cellular stress reflects mirna complementarity and correlates with altered translation rates. *Genes & development*, 27(14):1624–1632.

Kataoka, K., Shiraishi, Y., Takeda, Y., Sakata, S., Matsumoto, M., Nagano, S., Maeda, T., Nagata, Y., Kitanaka, A., Mizuno, S., et al. (2016). Aberrant PD-L1 expression through 3-UTR disruption in multiple cancers. *Nature*, 534(7607):402–406.

Kaymaz, Y., Oduor, C. I., Yu, H., Otieno, J. A., Ong'echa, J. M., Moormann, A. M., and Bailey, J. A. (2017). Comprehensive transcriptome and mutational profiling of endemic Burkitt lymphoma reveals EBV type–specific differences. *Molecular Cancer Research*, 15(5):563–576.

Kearse, M. G. and Wilusz, J. E. (2017). Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes & Development*, 31(17):1717–1731.

Kellaris, G., Khan, K., Baig, S. M., Tsai, I.-C., Zamora, F. M., Ruggieri, P., Natowicz, M. R., and Katsanis, N. (2018). A hypomorphic inherited pathogenic variant in DDX3X causes male intellectual disability with additional neurodevelopmental and neurodegenerative features. *Human Genomics*, 12(1):1–9.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome research*, 12(6):996–1006.

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, 26(17):2204–2207.

Ketteler, R. (2012). On programmed ribosomal frameshifting: the alternative proteomes. *Frontiers in genetics*, 3:242.

Kevil, C. G., De Benedetti, A., Payne, D. K., Coe, L. L., Laroux, F. S., and Alexander, J. S. (1996). Translational regulation of vascular permeability factor by eukaryotic initiation factor 4E: implications for tumor angiogenesis. *International Journal of Cancer*, 65(6):785–790.

Khodadoust, M. S., Olsson, N., Chen, B., Sworder, B., Shree, T., Liu, C. L., Zhang, L., Czerwinski, D. K., Davis, M. M., Levy, R., et al. (2019). B-cell lymphomas present immunoglobulin neoantigens. *Blood*, 133(8):878–881.

Khodadoust, M. S., Olsson, N., Wagar, L. E., Haabeth, O. A., Chen, B., Swaminathan, K., Rawson, K., Liu, C. L., Steiner, D., Lund, P., et al. (2017). Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. *Nature*, 543(7647):723–727.

Kim, J. and Guan, K. L. (2019a). mTOR as a central hub of nutrient signalling and cell growth. *Nature Cell Biology*, 21(1):63–71.

Kim, J. and Guan, K.-L. (2019b). mTOR as a central hub of nutrient signalling and cell growth. *Nature Cell Biology*, 21(1):63–71.

Kiyasu, J., Miyoshi, H., Hirata, A., Arakawa, F., Ichikawa, A., Niino, D., Sugita, Y., Yufu, Y., Choi, I., Abe, Y., et al. (2015). Expression of programmed cell death ligand 1 is associated with poor overall survival in patients with diffuse large b-cell lymphoma. *Blood*, 126(19):2193–2201.

Klein, U. and Dalla-Favera, R. (2008). Germinal centres: role in B-cell physiology and malignancy. *Nature Reviews Immunology*, 8(1):22–33.

Knight, J. R., Garland, G., Poyry, T., Mead, E., Vlahov, N., Sfakianos, A., Grosso, S., De-Lima-Hedayioglu, F., Mallucci, G. R., Von Der Haar, T., Smales, C. M., Sansom, O. J., and Willis, A. E. (2020). Control of translation elongation in health and disease. *DMM Disease Models and Mechanisms*, 13(3).

Komar, A. A. and Hatzoglou, M. (2011). Cellular IRES-mediated translation: the war of ITAFs in pathophysiological states. *Cell Cycle*, 10(2):229–240.

Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F., and Kageyama, Y. (2010). Small peptides switch the transcriptional activity of shavenbaby during drosophila embryogenesis. *Science*, 329(5989):336–339.

Kozak, M. (1987). At least six nucleotides preceding the aug initiator codon enhance translation in mammalian cells. *Journal of Molecular Biology*, 196(4):947–950.

Krokhin, O. V. and Spicer, V. (2010). Predicting peptide retention times for proteomics. *Current Protocols in Bioinformatics*, 31(1):13–14.

Kumari, P. and Sampath, K. (2015). cncRNAs: Bi-functional RNAs with protein coding and non-coding functions. In *Seminars in Cell & Developmental Biology*, volume 47, pages 40–51. Elsevier.

Küppers, R. and Dalla-Favera, R. (2001). Mechanisms of chromosomal translocations in B cell lymphomas. *Oncogene*, 20(40):5580–5594.

Kurosaki, T. and Maquat, L. E. (2016). Nonsense-mediated mRNA decay in humans at a glance. *Journal of Cell Science*, 129(3):461–467.

Kustatscher, G., Grabowski, P., Schrader, T. A., Passmore, J. B., Schrader, M., and Rappsilber, J. (2019). Co-regulation map of the human proteome enables identification of protein functions. *Nature Biotechnology*, 37(11):1361–1371.

Labun, K., Montague, T. G., Krause, M., Torres Cleuren, Y. N., Tjeldnes, H., and Valen, E. (2019). CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Research*, 47(W1):W171–W174.

Lacy, S. E., Barrans, S. L., Beer, P. A., Painter, D., Smith, A. G., Roman, E., Cooke, S. L., Ruiz, C., Glover, P., Van Hoppe, S. J., et al. (2020). Targeted sequencing in DLBCL, molecular subtypes, and outcomes: a haematological malignancy research network report. *Blood*, 135(20):1759–1771.

Lafontaine, D. L. (2015). Noncoding RNAs in eukaryotic ribosome biogenesis and function. *Nature Structural & Molecular Biology*, 22(1):11–19.

Lafontaine, D. L., Riback, J. A., Bascetin, R., and Brangwynne, C. P. (2021). The nucleolus as a multiphase liquid condensate. *Nature Reviews Molecular Cell Biology*, 22(3):165–182.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.

Larsson, O., Sonenberg, N., and Nadon, R. (2010). Identification of differential translation in genome wide studies. *Proceedings of the National Academy of Sciences*, 107(50):21487–21492.

Larsson, O., Sonenberg, N., and Nadon, R. (2011). anota: Analysis of differential translation in genome-wide studies. *Bioinformatics*, 27(10):1440–1441.

Lauria, F., Tebaldi, T., Bernabò, P., Groen, E. J., Gillingwater, T. H., and Viero, G. (2018). ribowaltz: optimization of ribosome p-site positioning in ribosome profiling data. *PLoS computational biology*, 14(8):e1006169.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):1–17.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9(8):e1003118.

Lawrie, C. H., Chi, J., Taylor, S., Tramonti, D., Ballabio, E., Palazzo, S., Saunders, N. J., Pezzella, F., Boultwood, J., Wainscoat, J. S., et al. (2009). Expression of microRNAs in diffuse large b cell lymphoma is associated with immunophenotype, survival and transformation from follicular lymphoma. *Journal of Cellular and Molecular Medicine*, 13(7):1248–1260.

Lazaris-Karatzas, A., Montine, K. S., and Sonenberg, N. (1990). Malignant transformation by a eukaryotic initiation factor subunit that binds to mRNA 5'cap. *Nature*, 345(6275):544–547.

Lee, A. S., Kranzusch, P. J., Doudna, J. A., and Cate, J. H. (2016). eIF3d is an mRNA cap-binding protein that is required for specialized translation initiation. *Nature*, 536(7614):96–99.

Lee, C.-S., Dias, A. P., Jedrychowski, M., Patel, A. H., Hsu, J. L., and Reed, R. (2008a). Human DDX3 functions in translation and interacts with the translation initiation factor eIF3. *Nucleic Acids Research*, 36(14):4708–4718.

Lee, C.-S., Dias, A. P., Jedrychowski, M., Patel, A. H., Hsu, J. L., and Reed, R. (2008b). Human DDX3 functions in translation and interacts with the translation initiation factor eIF3. *Nucleic Acids Research*, 36(14):4708–4718.

Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B., and Qian, S.-B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*, 109(37):E2424–E2432.

Lennox, A. L., Hoye, M. L., Jiang, R., Johnson-Kerner, B. L., Suit, L. A., Venkataramanan, S., Sheehan, C. J., Alsina, F. C., Fregeau, B., Aldinger, K. A., et al. (2020a). Pathogenic DDX3X mutations impair RNA metabolism and neurogenesis during fetal cortical development. *Neuron*, 106(3):404–420.

Lennox, A. L., Hoye, M. L., Jiang, R., Johnson-Kerner, B. L., Suit, L. A., Venkataramanan, S., Sheehan, C. J., Alsina, F. C., Fregeau, B., Aldinger, K. A., et al. (2020b). Pathogenic DDX3X mutations impair RNA metabolism and neurogenesis during fetal cortical development. *Neuron*, 106(3):404–420.

Leprivier, G., Remke, M., Rotblat, B., Dubuc, A., Mateo, A.-R. F., Kool, M., Agnihotri, S., El-Naggar, A., Yu, B., Somasekharan, S. P., et al. (2013). The eEF2 kinase confers resistance to nutrient deprivation by blocking translation elongation. *Cell*, 153(5):1064–1079.

Leucci, E., Cocco, M., Onnis, A., De Falco, G., Van Cleef, P., Bellan, C., Van Rijk, A., Nyagol, J., Byakika, B., Lazzi, S., et al. (2008). MYC translocation-negative classical Burkitt lymphoma cases: an alternative pathogenetic mechanism involving mirna deregulation. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 216(4):440–450.

Li, C., Kim, S.-W., Rai, D., Bolla, A. R., Adhvaryu, S., Kinney, M. C., Robetorye, R. S., and Aguiar, R. C. (2009a). Copy number abnormalities, MYC activity, and the genetic fingerprint of normal B cells mechanistically define the microRNA profile of diffuse large B-cell lymphoma. *Blood*, 113(26):6681–6690.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009b). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, J. and Liu, C. (2019). Coding or noncoding, the converging concepts of RNAs. *Frontiers in genetics*, 10:496.

Li, W., Wang, W., Uren, P. J., Penalva, L. O., and Smith, A. D. (2017). Riborex: fast and flexible identification of differential translation from Ribo-seq data. *Bioinformatics*, 33(11):1735–1737.

Liao, J.-M., Zhou, X., Gatignol, A., and Lu, H. (2014). Ribosomal proteins L5 and L11 co-operatively inactivate c-Myc via RNA-induced silencing complex. *Oncogene*, 33(41):4916–4923.

Linder, P. and Jankowsky, E. (2011). From unwinding to clamping the DEAD box RNA helicase family. *Nature Reviews Molecular Cell Biology*, 12(8):505–516.

Lindström, M. S., Jurada, D., Bursac, S., Orsolic, I., Bartek, J., and Volarevic, S. (2018). Nucleolus as an emerging hub in maintenance of genome stability and cancer pathogenesis. *Oncogene*, 37(18):2351–2366.

Liu, G. Y. and Sabatini, D. M. (2020). mTOR at the nexus of nutrition, growth, ageing and disease. *Nature Reviews Molecular Cell Biology*, 21(4):183–203.

Liu, J., Xu, Y., Stoleru, D., and Salic, A. (2012). Imaging protein synthesis in cells and tissues with an alkyne analog of puromycin. *Proceedings of the National Academy of Sciences*, 109(2):413–418.

Liu, P., Ge, M., Hu, J., Li, X., Che, L., Sun, K., Cheng, L., Huang, Y., Pilo, M. G., Cigliano, A., et al. (2017). A functional mammalian target of rapamycin complex 1 signaling is indispensable for c-Myc-driven hepatocarcinogenesis. *Hepatology*, 66(1):167–181.

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. *Cell*, 165(3):535–550.

López, C., Kleinheinz, K., Aukema, S. M., Rohde, M., Bernhart, S. H., Hübschmann, D., Wagener, R., Toprak, U. H., Raimondi, F., Kreuz, M., et al. (2019). Genomic and transcriptomic changes complement each other in the pathogenesis of sporadic Burkitt lymphoma. *Nature communications*, 10(1):1–19.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21.

Lu, J., Cannizzaro, E., Meier-Abt, F., Scheinost, S., Bruch, P.-M., Giles, H. A., Lütge, A., Hüllein, J., Wagner, L., Giacopelli, B., et al. (2021). Multi-omics reveals clinically relevant proliferative drive associated with mTOR-MYC-OXPHOS activity in chronic lymphocytic leukemia. *Nature Cancer*, 2(8):853–864.

Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Älgenäs, C., Lundeberg, J., Mann, M., and Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. *Molecular Systems Biology*, 6(1):450.

Lynch, M. and Marinov, G. K. (2015). The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences*, 112(51):15690–15695.

Mackowiak, S. D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., Mastrobuoni, G., Rajewsky, N., Kempa, S., Selbach, M., et al. (2015). Extensive identification and analysis of conserved small ORFs in animals. *Genome Biology*, 16(1):1–21.

Malumbres, R., Sarosiek, K. A., Cubedo, E., Ruiz, J. W., Jiang, X., Gascoyne, R. D., Tibshirani, R., and Lossos, I. S. (2009). Differentiation stage–specific expression of microRNAs in B lymphocytes and diffuse large B-cell lymphomas. *Blood*, 113(16):3754–3764.

Mangus, D. A., Evans, M. C., and Jacobson, A. (2003). Poly (A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome Biology*, 4(7):1–14.

Maquat, L. E., Tarn, W.-Y., and Isken, O. (2010). The pioneer round of translation: features and functions. *Cell*, 142(3):368–374.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, 17(1):10–12.

Mathieson, T., Franken, H., Kosinski, J., Kurzawa, N., Zinn, N., Sweetman, G., Poeckel, D., Ratnu, V. S., Schramm, M., Becher, I., et al. (2018). Systematic analysis of protein turnover in primary cells. *Nature Communications*, 9(1):1–10.

McCord, R., Bolen, C. R., Koeppen, H., Kadel, E. E., Oestergaard, M. Z., Nielsen, T., Sehn, L. H., and Venstrom, J. M. (2019a). PD-L1 and tumor-associated macrophages in de novo DLBCL. *Blood Advances*, 3(4):531–540.

McCord, R., Bolen, C. R., Koeppen, H., Kadel, E. E., Oestergaard, M. Z., Nielsen, T., Sehn, L. H., and Venstrom, J. M. (2019b). PD-L1 and tumor-associated macrophages in de novo DLBCL. *Blood Advances*, 3(4):531–540.

McGlincy, N. J. and Ingolia, N. T. (2017). Transcriptome-wide measurement of translation by ribosome profiling. *Methods*, 126:112–129.

McMahon, S. B. (2014). MYC and the control of apoptosis. *Cold Spring Harbor perspectives in medicine*, 4(7):a014407.

Meyer, K. D., Patil, D. P., Zhou, J., Zinoviev, A., Skabkin, M. A., Elemento, O., Pestova, T. V., Qian, S.-B., and Jaffrey, S. R. (2015). 5 UTR m6A promotes cap-independent translation. *Cell*, 163(4):999–1010.

Meyer, N. and Penn, L. Z. (2008). Reflecting on 25 years with MYC. *Nature Reviews Cancer*, 8(12):976–990.

Mlynarczyk, C., Fontán, L., and Melnick, A. (2019). Germinal center-derived lymphomas: The darkest side of humoral immunity. *Immunological Reviews*, 288(1):214–239.

Mo, J., Liang, H., Su, C., Li, P., Chen, J., and Zhang, B. (2021). DDX3X: structure, physiologic functions and cancer. *Molecular cancer*, 20(1):1–20.

Modelska, A., Turro, E., Russell, R., Beaton, J., Sbarrato, T., Spriggs, K., Miller, J., Gräf, S., Provenzano, E., Blows, F., et al. (2015). The malignant phenotype in breast cancer is driven by eif4a1-mediated changes in the translational landscape. *Cell death & disease*, 6(1):e1603–e1603.

Mohr, I. (2016). Virology: Closing in on the causes of host shutoff. *eLife*, 5:e20755.

Molliex, A., Temirov, J., Lee, J., Coughlin, M., Kanagaraj, A. P., Kim, H. J., Mittag, T., and Taylor, J. P. (2015). Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell*, 163(1):123–133.

Moreno-Mateos, M. A., Vejnar, C. E., Beaudoin, J.-D., Fernandez, J. P., Mis, E. K., Khokha, M. K., and Giraldez, A. J. (2015). CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nature Methods*, 12(10):982–988.

Morgan, M., Pagès, H., Obenchain, V., and Hayden, N. (2016). Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. *R package version*, 1(0):677–689.

Morrish, F. and Hockenbery, D. (2014). MYC and mitochondrial biogenesis. *Cold Spring Harbor perspectives in medicine*, 4(5):a014225.

Morton, L. M., Wang, S. S., Devesa, S. S., Hartge, P., Weisenburger, D. D., and Linet, M. S. (2006). Lymphoma incidence patterns by WHO subtype in the United States, 1992-2001. *Blood*, 107(1):265–276.

Moser, B. and Willimann, K. (2004). Chemokines: role in inflammation and immune surveillance. *Annals of the rheumatic diseases*, 63(suppl 2):ii84–ii89.

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell*, 102(5):553–563.

Nakagawa, R. and Calado, D. P. (2021). Positive selection in the light zone of germinal centers. *Frontiers in Immunology*, 12:1053.

Nascimento, E. M., Cox, C. L., MacArthur, S., Hussain, S., Trotter, M., Blanco, S., Suraj, M., Nichols, J., Kübler, B., Benitah, S. A., et al. (2011). The opposing transcriptional functions of Sin3a and c-Myc are required to maintain tissue homeostasis. *Nature Cell Biology*, 13(12):1395–1405.

Navarro, A., Gaya, A., Martinez, A., Urbano-Ispizua, A., Pons, A., Balagué, O., Gel, B., Abrisqueta, P., Lopez-Guillermo, A., Artells, R., et al. (2008). MicroRNA expression profiling in classic Hodgkin lymphoma. *Blood*, 111(5):2825–2832.

Neelagandan, N., Lamberti, I., Carvalho, H. J., Gobet, C., and Naef, F. (2020). What determines eukaryotic translation elongation: recent molecular and quantitative analyses of protein synthesis. *Open biology*, 10(12):200292.

Nesvizhskii, A. I. (2014). Proteogenomics: concepts, applications and computational strategies. *Nature Methods*, 11(11):1114–1125.

Northcott, P. A., Buchhalter, I., Morrissy, A. S., Hovestadt, V., Weischenfeldt, J., Ehrenberger, T., Gröbner, S., Segura-Wang, M., Zichner, T., Rudneva, V. A., et al. (2017). The whole-genome landscape of medulloblastoma subtypes. *Nature*, 547(7663):311–317.

Obrig, T. G., Culp, W. J., McKeehan, W. L., and Hardesty, B. (1971). The mechanism by which cycloheximide and related glutarimide antibiotics inhibit peptide synthesis on reticulocyte ribosomes. *Journal of Biological Chemistry*, 246(1):174–181.

Oertlin, C., Lorent, J., Murie, C., Furic, L., Topisirovic, I., and Larsson, O. (2019). Generally applicable transcriptome-wide analysis of translation using anota2seq. *Nucleic Acids Research*, 47(12):e70–e70.

Oh, S., Flynn, R. A., Floor, S. N., Purzner, J., Martin, L., Do, B. T., Schubert, S., Vaka, D., Morrissy, S., Li, Y., et al. (2016). Medulloblastoma-associated DDX3 variant selectively alters the translational response to stress. *Oncotarget*, 7(19):28169.

Ojha, J., Ayres, J., Secreto, C., Tschumper, R., Rabe, K., Van Dyke, D., Slager, S., Shanafelt, T., Fonseca, R., Kay, N. E., et al. (2015). Deep sequencing identifies genetic heterogeneity and recurrent convergent evolution in chronic lymphocytic leukemia. *Blood, The Journal of the American Society of Hematology*, 125(3):492–498.

Olexiouk, V., Van Criekinge, W., and Menschaert, G. (2018). An update on sORFs. org: a repository of small orfs identified by ribosome profiling. *Nucleic Acids Research*, 46(D1):D497–D502.

Orr, M. W., Mao, Y., Storz, G., and Qian, S.-B. (2020). Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Research*, 48(3):1029–1042.

Ouspenskaia, T., Law, T., Clauser, K. R., Klaeger, S., Sarkizova, S., Aguet, F., Li, B., Christian, E., Knisbacher, B. A., Le, P. M., Hartigan, C. R., Keshishian, H., Apffel, A., Oliveira, G., Zhang, W., Chow, Y. T., Ji, Z., Shukla, S. A., Bachireddy, P., Getz, G., Hacohen, N., Keskin, D. B., Carr, S. A., Wu, C. J., and Regev, A. (2020). Thousands of novel unannotated proteins expand the MHC I immunopeptidome in cancer. *bioRxiv*.

Ozuah, N. W., Lubega, J., Allen, C. E., and El-Mallawany, N. K. (2020). Five decades of low intensity and low survival: adapting intensified regimens to cure pediatric Burkitt lymphoma in Africa. *Blood Advances*, 4(16):4007–4019.

Pakos-Zebrucka, K., Koryga, I., Mnich, K., Ljujic, M., Samali, A., and Gorman, A. M. (2016). The integrated stress response. *EMBO reports*, 17(10):1374–1395.

Palade, G. E. (1955). A small particulate component of the cytoplasm. *The Journal of Cell Biology*, 1(1):59–68.

Pardoll, D. M. (2012). The blockade of immune checkpoints in cancer immunotherapy. *Nature Reviews Cancer*, 12(4):252–264.

Patmore, D. M., Jassim, A., Nathan, E., Gilbertson, R. J., Tahan, D., Hoffmann, N., Tong, Y., Smith, K. S., Kanneganti, T.-D., Suzuki, H., et al. (2020). DDX3X suppresses the susceptibility of hindbrain lineages to medulloblastoma. *Developmental Cell*, 54(4):455–470.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417–419.

Pearson, H., Daouda, T., Granados, D. P., Durette, C., Bonneil, E., Courcelles, M., Rodenbrock, A., Laverdure, J.-P., Côté, C., Mader, S., et al. (2016). MHC class I–associated peptides derive from selective regions of the human genome. *The Journal of clinical investigation*, 126(12):4690–4701.

Pearson, K. (1897). On a form of spurious correlation which may arise when indices are used in the measurement of organs. In *Royal Soc., London, Proc.*, volume 60, pages 489–502.

Pelletier, J., Thomas, G., and Volarević, S. (2018). Ribosome biogenesis in cancer: new players and therapeutic avenues. *Nature Reviews Cancer*, 18(1):51–63.

Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., Signal, B., Gloss, B. S., Hammang, C. J., Rost, B., et al. (2015). Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences*, 112(52):15898–15903.

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., et al. (2019). The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research*, 47(D1):D442–D450.

Perez-Riverol, Y., Vizcaíno, J. A., and Griss, J. (2018). Future prospects of spectral clustering approaches in proteomics. *Proteomics*, 18(14):1700454.

Pfeifer, M., Grau, M., Lenze, D., Wenzel, S.-S., Wolf, A., Wollert-Wulf, B., Dietze, K., Nogai, H., Storek, B., Madle, H., et al. (2013). PTEN loss defines a PI3K/AKT pathway-dependent germinal center subtype of diffuse large B-cell lymphoma. *Proceedings of the National Academy of Sciences*, 110(30):12420–12425.

Pfister, A. S. (2019). Emerging role of the nucleolar stress response in autophagy. *Frontiers in Cellular Neuroscience*, 13:156.

Phelan, J. D., Young, R. M., Webster, D. E., Roulland, S., Wright, G. W., Kasbekar, M., Shaffer, A. L., Ceribelli, M., Wang, J. Q., Schmitz, R., et al. (2018). A multiprotein supercomplex controlling oncogenic signalling in lymphoma. *Nature*, 560(7718):387–391.

Philippe, L., van den Elzen, A. M., Watson, M. J., and Thoreen, C. C. (2020). Global analysis of LARP1 translation targets reveals tunable and dynamic features of 5TOP motifs. *Proceedings of the National Academy of Sciences*, 117(10):5319–5328.

Phung, B., Cieśla, M., Sanna, A., Guzzi, N., Beneventi, G., Ngoc, P. C. T., Lauss, M., Cabrita, R., Cordero, E., Bosch, A., et al. (2019). The X-linked DDX3X RNA helicase dictates translation reprogramming and metastasis in melanoma. *Cell Reports*, 27(12):3573–3586.

Pianese, G. (1896). *Beitrag zur histologie und aetiologie des carcinoms*, volume 1. G. Fischer.

Piccirillo, C. A., Bjur, E., Topisirovic, I., Sonenberg, N., and Larsson, O. (2014). Translational control of immune responses: from transcripts to translatomes. *Nature Immunology*, 15(6):503–511.

Pourdehnad, M., Truitt, M. L., Siddiqi, I. N., Ducker, G. S., Shokat, K. M., and Ruggero, D. (2013). Myc and mTOR converge on a common node in protein synthesis control that confers synthetic lethality in Myc-driven cancers. *Proceedings of the National Academy of Sciences*, 110(29):11988–11993.

Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K. E., Graveley, B. R., et al. (2015). Codon optimality is a major determinant of mRNA stability. *Cell*, 160(6):1111–1124.

Protter, D. S. and Parker, R. (2016). Principles and properties of stress granules. *Trends in Cell Biology*, 26(9):668–679.

Pugh, T. J., Weeraratne, S. D., Archer, T. C., Krummel, D. A. P., Auclair, D., Bochicchio, J., Carneiro, M. O., Carter, S. L., Cibulskis, K., Erlich, R. L., et al. (2012). Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature*, 488(7409):106–110.

Pyronnet, S., Imataka, H., Gingras, A.-C., Fukunaga, R., Hunter, T., and Sonenberg, N. (1999). Human eukaryotic translation initiation factor 4G (eIF4G) recruits mnk1 to phosphorylate eIF4E. *The EMBO journal*, 18(1):270–279.

Rakhra, K., Bachireddy, P., Zabuawala, T., Zeiser, R., Xu, L., Kopelman, A., Fan, A. C., Yang, Q., Braunstein, L., Crosby, E., et al. (2010). CD4+ T cells contribute to the remodeling of the microenvironment required for sustained tumor regression upon oncogene inactivation. *Cancer Cell*, 18(5):485–498.

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., and Manke, T. (2016). deeptools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1):W160–W165.

Ratje, A. H., Loerke, J., Mikolajka, A., Brünner, M., Hildebrand, P. W., Starosta, A. L., Dönhöfer, A., Connell, S. R., Fucini, P., Mielke, T., et al. (2010). Head swivel on the ribosome facilitates translocation by means of intra-subunit tRNA hybrid sites. *Nature*, 468(7324):713–716.

Raught, B. and Gingras, A.-C. (1999). eIF4E activity is regulated at multiple levels. *The international journal of biochemistry & cell biology*, 31(1):43–57.

Rauschendorf, M.-A., Zimmer, J., Hanstein, R., Dickemann, C., and Vogt, P. (2011). Complex transcriptional control of the AZFa gene DDX3Y in human testis. *International Journal of Andrology*, 34(1):84–96.

Reddy, A., Zhang, J., Davis, N. S., Moffitt, A. B., Love, C. L., Waldrop, A., Leppa, S., Pasanen, A., Meriranta, L., Karjalainen-Lindsberg, M.-L., et al. (2017). Genetic and functional drivers of diffuse large b cell lymphoma. *Cell*, 171(2):481–494.

Richter, J., Schlesner, M., Hoffmann, S., Kreuz, M., Leich, E., Burkhardt, B., Rosolowski, M., Ammerpohl, O., Wagener, R., Bernhart, S. H., et al. (2012). Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nature Genetics*, 44(12):1316.

Richter, J. D. and Coller, J. (2015). Pausing on polyribosomes: make way for elongation in translational control. *Cell*, 163(2):292–300.

Robichaud, N., Hsu, B. E., Istomine, R., Alvarez, F., Blagih, J., Ma, E. H., Morales, S. V., Dai, D. L., Li, G., Souleimanova, M., et al. (2018). Translational control in the tumor microenvironment promotes lung metastasis: Phosphorylation of eif4e in neutrophils. *Proceedings of the National Academy of Sciences*, 115(10):E2202–E2209.

Robinson, G., Parker, M., Kranenburg, T. A., Lu, C., Chen, X., Ding, L., Phoenix, T. N., Hedlund, E., Wei, L., Zhu, X., et al. (2012). Novel mutations target distinct subgroups of medulloblastoma. *Nature*, 488(7409):43–48.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Rogozin, I. B., Kochetov, A. V., Kondrashov, F. A., Koonin, E. V., and Milanesi, L. (2001). Presence of ATG triplets in 5 untranslated regions of eukaryotic cDNAs correlates with a 'weak'context of the start codon. *Bioinformatics*, 17(10):890–900.

Rolfe, D. and Brown, G. C. (1997). Cellular energy utilization and molecular origin of standard metabolic rate in mammals. *Physiological Reviews*, 77(3):731–758.

Ron, D. and Walter, P. (2007). Signal integration in the endoplasmic reticulum unfolded protein response. *Nature Reviews Molecular cell biology*, 8(7):519–529.

Rooijers, K., Loayza-Puch, F., Nijtmans, L. G., and Agami, R. (2013). Ribosome profiling reveals features of normal and disease-associated mitochondrial translation. *Nature Communications*, 4(1):1–8.

Rouschop, K. M., Van Den Beucken, T., Dubois, L., Niessen, H., Bussink, J., Savelkouls, K., Keulers, T., Mujcic, H., Landuyt, W., Voncken, J. W., et al. (2010). The unfolded protein response protects human tumor cells during hypoxia through regulation of the autophagy genes MAP1LC3B and ATG5. *The Journal of Clinical Investigation*, 120(1):127–141.

Ruggero, D. (2013). Translational control in cancer etiology. *Cold Spring Harbor Perspectives in Biology*, 5(2):a012336.

Ruggero, D., Montanaro, L., Ma, L., Xu, W., Londei, P., Cordon-Cardo, C., and Pandolfi, P. P. (2004). The translation factor eIF-4E promotes tumor formation and cooperates with c-Myc in lymphomagenesis. *Nature medicine*, 10(5):484–486.

Ruggiano, A., Foresti, O., and Carvalho, P. (2014). Er-associated degradation: Protein quality control and beyond. *Journal of Cell Biology*, 204(6):869–879.

Ruiz-Orera, J., Messeguer, X., Subirana, J. A., and Alba, M. M. (2014). Long non-coding RNAs as a source of new peptides. *eLife*, 3:e03523.

Runte, F., Renner IV, P., and Hoppe, M. (2019). Kuby immunology.

Sabi, R. and Tuller, T. (2014). Modelling the Efficiency of Codon–tRNA Interactions Based on Codon Usage Bias. *DNA Research*, 21(5):511–526.

Sadedin, S. P., Pope, B., and Oshlack, A. (2012). Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics*, 28(11):1525–1526.

Samir, P., Kesavardhana, S., Patmore, D. M., Gingras, S., Malireddi, R. S., Karki, R., Guy, C. S., Briard, B., Place, D. E., Bhattacharya, A., et al. (2019). DDX3X acts as a live-or-die checkpoint in stressed cells by regulating NLRP3 inflammasome. *Nature*, 573(7775):590–594.

Sander, S., Calado, D. P., Srinivasan, L., Köchert, K., Zhang, B., Rosolowski, M., Rodig, S. J., Holzmann, K., Stilgenbauer, S., Siebert, R., et al. (2012). Synergy between PI3K signaling and MYC in Burkitt lymphomagenesis. *Cancer Cell*, 22(2):167–179.

Sanson, K. R., Hanna, R. E., Hegde, M., Donovan, K. F., Strand, C., Sullender, M. E., Vaimberg, E. W., Goodale, A., Root, D. E., Piccioni, F., et al. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nature Communications*, 9(1):1–15.

Santos, D. A., Shi, L., Tu, B. P., and Weissman, J. S. (2019). Cycloheximide can distort measurements of mRNA levels and translation efficiency. *Nucleic acids research*, 47(10):4974–4985.

Sanz, E., Yang, L., Su, T., Morris, D. R., McKnight, G. S., and Amieux, P. S. (2009). Cell-type-specific isolation of ribosome-associated mRNA from complex tissues. *Proceedings of the National Academy of Sciences*, 106(33):13939–13944.

Sarkizova, S., Klaeger, S., Le, P. M., Li, L. W., Oliveira, G., Keshishian, H., Hartigan, C. R., Zhang, W., Braun, D. A., Ligon, K. L., et al. (2020). A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nature Biotechnology*, 38(2):199–209.

Saxton, R. A. and Sabatini, D. M. (2017). mTOR Signaling in Growth, Metabolism, and Disease. *Cell*, 168(6):960–976.

Schatz, J. H., Oricchio, E., Wolfe, A. L., Jiang, M., Linkov, I., Maragulia, J., Shi, W., Zhang, Z., Rajasekhar, V. K., Pagano, N. C., et al. (2011). Targeting cap-dependent translation blocks converging survival signals by AKT and PIM kinases in lymphoma. *Journal of Experimental Medicine*, 208(9):1799–1807.

Schmidt, E. V. (2004). The role of c-myc in regulation of translation initiation. *Oncogene*, 23(18):3217–3221.

Schmitz, R., Wright, G. W., Huang, D. W., Johnson, C. A., Phelan, J. D., Wang, J. Q., Roulland, S., Kasbekar, M., Young, R. M., Shaffer, A. L., et al. (2018). Genetics and pathogenesis of diffuse large b-cell lymphoma. *New England Journal of Medicine*, 378(15):1396–1407.

Schmitz, R., Young, R. M., Ceribelli, M., Jhavar, S., Xiao, W., Zhang, M., Wright, G., Shaffer, A. L., Hodson, D. J., Buras, E., et al. (2012). Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature*, 490(7418):116–120.

Schneider-Poetsch, T., Ju, J., Eyler, D. E., Dang, Y., Bhat, S., Merrick, W. C., Green, R., Shen, B., and Liu, J. O. (2010). Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nature Chemical Biology*, 6(3):209–217.

Schueren, F. and Thoms, S. (2016). Functional translational readthrough: a systems biology perspective. *PLoS Genetics*, 12(8):e1006196.

Schuller, A. P. and Green, R. (2018). Roadblocks and resolutions in eukaryotic translation. *Nature Reviews Molecular Cell Biology*, 19(8):526–541.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342.

Sen, N. D., Zhou, F., Ingolia, N. T., and Hinnebusch, A. G. (2015). Genome-wide analysis of translational efficiency reveals distinct but overlapping functions of yeast DEAD-box RNA helicases Ded1 and eIF4A. *Genome Research*, 25(8):1196–1205.

Sendoel, A., Dunn, J. G., Rodriguez, E. H., Naik, S., Gomez, N. C., Hurwitz, B., Levorse, J., Dill, B. D., Schramek, D., Molina, H., et al. (2017). Translation from unconventional 5 start sites drives tumour initiation. *Nature*, 541(7638):494–499.

Sha, C., Barrans, S., Cucco, F., Bentley, M. A., Care, M. A., Cummin, T., Kennedy, H., Thompson, J. S., Uddin, R., Worrillow, L., et al. (2019). Molecular high-grade b-cell lymphoma: defining a poor-risk group that requires different approaches to therapy. *Journal of Clinical Oncology*, 37(3):202.

Shaffer III, A. L., Young, R. M., and Staudt, L. M. (2012). Pathogenesis of human b cell lymphomas. *Annual review of immunology*, 30:565–610.

Sharma, P., Nilges, B. S., Wu, J., and Leidel, S. A. (2019). The translation inhibitor cycloheximide affects ribosome profiling data in a species-specific manner. *bioRxiv*, page 746255.

Sharpe, A. H. and Pauken, K. E. (2018). The diverse functions of the PD1 inhibitory pathway. *Nature Reviews Immunology*, 18(3):153–167.

Shatsky, I. N., Terenin, I. M., Smirnova, V. V., and Andreev, D. E. (2018). Cap-independent translation: What's in a name? *Trends in Biochemical Sciences*, 43(11):882–895.

Shi, Z. and Barna, M. (2015). Translating the genome in time and space: specialized ribosomes, RNA regulons, and RNA-binding proteins. *Annual Review of Cell and Developmental Biology*, 31:31–54.

Shi, Z., Fujii, K., Kovary, K. M., Genuth, N. R., Röst, H. L., Teruel, M. N., and Barna, M. (2017). Heterogeneous ribosomes preferentially translate distinct subpools of mRNAs genome-wide. *Molecular cell*, 67(1):71–83.

Shih, J., Tsai, T., Chao, C.-H., and Lee, Y. W. (2008). Candidate tumor suppressor DDX3 RNA helicase specifically represses cap-dependent translation by acting as an eIF4E inhibitory protein. *Oncogene*, 27(5):700–714.

Shiue, C. N., Berkson, R. G., and Wright, A. P. (2009). c-Myc induces changes in higher order rDNA structure on stimulation of quiescent cells. *Oncogene*, 28(16):1833–1842.

Silvera, D., Formenti, S. C., and Schneider, R. J. (2010). Translational control in cancer. *Nature Reviews Cancer*, 10(4):254–266.

Skourti-Stathaki, K. and Proudfoot, N. J. (2014). A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes & Development*, 28(13):1384–1396.

Smith, A., Crouch, S., Lax, S., Li, J., Painter, D., Howell, D., Patmore, R., Jack, A., and Roman, E. (2015). Lymphoma incidence, survival and prevalence 2004-2014: Sub-type analyses from the UK's Haematological Malignancy Research Network. *British Journal of Cancer*, 112(9):1575–1584.

Smith, L. M. and Kelleher, N. L. (2013). Proteoform: a single term describing protein complexity. *Nature methods*, 10(3):186–187.

Smith, R. C., Kanellos, G., Vlahov, N., Alexandrou, C., Willis, A. E., Knight, J. R., and Sansom, O. J. (2021). Translation initiation in cancer at a glance. *Journal of Cell Science*, 134(1).

Sole, C., Larrea, E., Manterola, L., Goicoechea, I., Armesto, M., Arestin, M., M Caffarel, M., M Araujo, A., Fernandez-Mercado, M., Araiz, M., et al. (2016). Aberrant expression of microRNAs in B-cell lymphomas. *Microrna*, 5(2):87–105.

Sonenberg, N. (1996). mRNA 5'cap-binding protein eIF4E and control of cell growth. *Translational Control*.

Sonenberg, N. and Hinnebusch, A. G. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, 136(4):731–745.

Song, S., Cao, C., Choukrallah, M.-A., Tang, F., Christofori, G., Kohler, H., Wu, F., Fodor, B. D., Frederiksen, M., Willis, S. N., et al. (2021). OBF1 and Oct factors control the germinal center transcriptional program. *Blood, The Journal of the American Society of Hematology*, 137(21):2920–2934.

Soto-Rifo, R., Rubilar, P. S., Limousin, T., De Breyne, S., Decimo, D., and Ohlmann, T. (2012). DEAD-box protein DDX3 associates with eif4f to promote translation of selected mRNAs. *The EMBO journal*, 31(18):3745–3756.

Staege, M. S., Lee, S. P., Frisan, T., Mautner, J., Scholz, S., Pajic, A., Rickinson, A. B., Masucci, M. G., Polack, A., and Bornkamm, G. W. (2002). MYC overexpression imposes a nonimmunogenic phenotype on Epstein–Barr virus-infected B cells. *Proceedings of the National Academy of Sciences*, 99(7):4550–4555.

Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology*, 22(2):96–118.

Steinhardt, J. J., Peroutka, R. J., Mazan-Mamczarz, K., Chen, Q., Houng, S., Robles, C., Barth, R. N., DuBose, J., Bruns, B., Tesoriero, R., et al. (2014). Inhibiting CARD11 translation during BCR activation by targeting the eIF4A RNA helicase. *Blood*, 124(25):3758–3767.

Stransky, N., Egloff, A. M., Tward, A. D., Kostic, A. D., Cibulskis, K., Sivachenko, A., Kryukov, G. V., Lawrence, M. S., Sougnez, C., McKenna, A., et al. (2011). The mutational landscape of head and neck squamous cell carcinoma. *Science*, 333(6046):1157–1160.

Stults, D. M., Killen, M. W., Williamson, E. P., Hourigan, J. S., Vargas, H. D., Arnold, S. M., Moscow, J. A., and Pierce, A. J. (2009). Human rRNA gene clusters are recombinational hotspots in cancer. *Cancer Research*, 69(23):9096–9104.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Suresh, S., Chen, B., Zhu, J., Golden, R. J., Lu, C., Evers, B. M., Novaresi, N., Smith, B., Zhan, X., Schmid, V., et al. (2020). eIF5B drives integrated stress response-dependent translation of PD-L1 in lung cancer. *Nature Cancer*, 1(5):533–545.

Swerdlow, S. H., Campo, E., Pileri, S. A., Harris, N. L., Stein, H., Siebert, R., Advani, R., Ghielmini, M., Salles, G. A., Zelenetz, A. D., et al. (2016). The 2016 revision of the world health organization classification of lymphoid neoplasms. *Blood*, 127(20):2375–2390.

Takahashi, K., Hu, B., Wang, F., Yan, Y., Kim, E., Vitale, C., Patel, K. P., Strati, P., Gumbs, C., Little, L., et al. (2018). Clinical implications of cancer gene mutations in patients with chronic lymphocytic leukemia treated with lenalidomide. *Blood*, 131(16):1820–1832.

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 47(D1):D941–D947.

Taylor, J., Yeomans, A. M., and Packham, G. (2020). Targeted inhibition of mRNA translation initiation factors as a novel therapeutic strategy for mature B-cell neoplasms. *Exploration of targeted anti-tumor therapy*, 1:3.

Terenin, I. M., Andreev, D. E., Dmitriev, S. E., and Shatsky, I. N. (2013). A novel mechanism of eukaryotic translation initiation that is neither m7G-cap-, nor IRES-dependent. *Nucleic Acids Research*, 41(3):1807–1816.

Thoreen, C. C., Chantranupong, L., Keys, H. R., Wang, T., Gray, N. S., and Sabatini, D. M. (2012). A unifying model for mTORC1-mediated regulation of mRNA translation. *Nature*, 485(7396):109–113.

Thyme, S. B., Akhmetova, L., Montague, T. G., Valen, E., and Schier, A. F. (2016). Internal guide RNA interactions interfere with Cas9-mediated cleavage. *Nature Communications*, 7(1):1–7.

Tompa, P., Davey, N. E., Gibson, T. J., and Babu, M. M. (2014). A million peptide motifs for the molecular biologist. *Molecular Cell*, 55(2):161–169.

Torrent, M., Chalancon, G., de Groot, N. S., Wuster, A., and Madan Babu, M. (2018). Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Science Signaling*, 11(546).

Truitt, M. L., Conn, C. S., Shi, Z., Pang, X., Tokuyasu, T., Coady, A. M., Seo, Y., Barna, M., and Ruggero, D. (2015). Differential requirements for eIF4E dose in normal development and cancer. *Cell*, 162(1):59–71.

Tschochner, H. and Hurt, E. (2003). Pre-ribosomes on the road from the nucleolus to the cytoplasm. *Trends in Cell Biology*, 13(5):255–263.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2):344–354.

Twa, D. D., Chan, F. C., Ben-Neriah, S., Woolcock, B. W., Mottok, A., Tan, K. L., Slack, G. W., Gunawardana, J., Lim, R. S., McPherson, A. W., et al. (2014). Genomic rearrangements involving programmed death ligands are recurrent in primary mediastinal large b-cell lymphoma. *Blood*, 123(13):2062–2065.

Unterluggauer, J. J., Prochazka, K., Tomazic, P. V., Huber, H. J., Seeboeck, R., Fechter, K., Steinbauer, E., Gruber, V., Feichtinger, J., Pichler, M., et al. (2018). Expression profile of translation initiation factor eIF2B5 in diffuse large B-cell lymphoma and its correlation to clinical outcome. *Blood Cancer Journal*, 8(9):1–5.

Urra, H., Dufey, E., Avril, T., Chevet, E., and Hetz, C. (2016). Endoplasmic reticulum stress and the hallmarks of cancer. *Trends in Cancer*, 2(5):252–262.

Valentin-Vega, Y. A., Wang, Y.-D., Parker, M., Patmore, D. M., Kanagaraj, A., Moore, J., Rusch, M., Finkelstein, D., Ellison, D. W., Gilbertson, R. J., et al. (2016). Cancer-associated DDX3X mutations drive stress granule assembly and impair global translation. *Scientific reports*, 6(1):1–16.

Van der Auwera, G. A. and O'Connor, B. D. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media.

van Heesch, S., Witte, F., Schneider-Lunitz, V., Schulz, J. F., Adami, E., Faber, A. B., Kirchner, M., Maatz, H., Blachut, S., Sandmann, C.-L., et al. (2019). The translational landscape of the human heart. *Cell*, 178(1):242–260.

Van Riggelen, J., Yetil, A., and Felsher, D. W. (2010). MYC as a regulator of ribosome biogenesis and protein synthesis. *Nature Reviews Cancer*, 10(4):301–309.

Van Steeg, H., Van Oostrom, C. T., Hodemaekers, H. M., Peters, L., and Thomas, A. A. (1991). The translation in vitro of rat ornithine decarboxylase mRNA is blocked by its 5 untranslated region in a polyamine-independent way. *Biochemical Journal*, 274(2):521–526.

Vattem, K. M. and Wek, R. C. (2004). Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proceedings of the National Academy of Sciences*, 101(31):11269–11274.

Venkataramanan, S., Calviello, L., Wilkins, K., and Floor, S. N. (2020). DDX3X and DDX3Y are redundant in protein synthesis. *Biorxiv*.

Versteeg, R., Noordermeer, I. A., Krüse-Wolters, M., Ruiter, D. J., and Schrier, P. I. (1988). c-myc down-regulates class I HLA expression in human melanomas. *The EMBO journal*, 7(4):1023–1029.

Walter, P. and Ron, D. (2011). The unfolded protein response: from stress pathway to homeostatic regulation. *Science*, 334(6059):1081–1086.

Wang, X., Zhao, B. S., Roundtree, I. A., Lu, Z., Han, D., Ma, H., Weng, X., Chen, K., Shi, H., and He, C. (2015). N6-methyladenosine modulates messenger RNA translation efficiency. *Cell*, 161(6):1388–1399.

Wei, J., Kishton, R. J., Angel, M., Conn, C. S., Dalla-Venezia, N., Marcel, V., Vincent, A., Catez, F., Ferré, S., Ayadi, L., et al. (2019). Ribosomal proteins regulate MHC class I peptide generation for immunosurveillance. *Molecular cell*, 73(6):1162–1173.

Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B., and Bartel, D. P. (2016). Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell Reports*, 14(7):1787–1799.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120.

Wendel, H.-G., De Stanchina, E., Fridman, J. S., Malina, A., Ray, S., Kogan, S., Cordon-Cardo, C., Pelletier, J., and Lowe, S. W. (2004). Survival signalling by Akt and eIF4E in oncogenesis and cancer therapy. *Nature*, 428(6980):332–337.

Wethmar, K., Barbosa-Silva, A., Andrade-Navarro, M. A., and Leutz, A. (2014). uORFdb — a comprehensive literature database on eukaryotic uorf biology. *Nucleic Acids Research*, 42(D1):D60–D67.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilmore, S., Rogers-Broadway, K.-R., Taylor, J., Lemm, E., Fell, R., Stevenson, F. K., Forconi, F., Steele, A. J., Coldwell, M., Packham, G., et al. (2021). Targeted inhibition of eIF4A suppresses b-cell receptor-induced translation and expression of MYC and MCL1 in chronic lymphocytic leukemia cells. *Cellular and Molecular Life Sciences*, pages 1–13.

Wolfe, A. L., Singh, K., Zhong, Y., Drewe, P., Rajasekhar, V. K., Sanghvi, V. R., Mavrakis, K. J., Jiang, M., Roderick, J. E., Van der Meulen, J., et al. (2014a). Rna g-quadruplexes cause eif4a-dependent oncogene translation in cancer. *Nature*, 513(7516):65–70.

Wolfe, A. L., Singh, K., Zhong, Y., Drewe, P., Rajasekhar, V. K., Sanghvi, V. R., Mavrakis, K. J., Jiang, M., Roderick, J. E., Van der Meulen, J., Schatz, J. H., Rodrigo, C. M., Zhao, C., Rondou, P., de Stanchina, E., Teruya-Feldstein, J., Kelliher, M. A., Speleman, F., Porco, J. A., Pelletier, J., Rätsch, G., and Wendel, H. G. (2014b). RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Nature*, 513(7516):65–70.

Wolin, S. L. and Walter, P. (1988). Ribosome pausing and stacking during translation of a eukaryotic mRNA. *The EMBO journal*, 7(11):3559–3569.

Xiang, N., He, M., Ishaq, M., Gao, Y., Song, F., Guo, L., Ma, L., Sun, G., Liu, D., Guo, D., et al. (2016). The DEAD-box RNA helicase DDX3 interacts with nf-κb subunit p65 and suppresses p65-mediated transcription. *PloS one*, 11(10):e0164471.

Xiao, Z., Huang, R., Xing, X., Chen, Y., Deng, H., and Yang, X. (2018). De novo annotation and characterization of the translatome with ribosome profiling data. *Nucleic Acids Research*, 46(10):e61–e61.

Xiao, Z., Zou, Q., Liu, Y., and Yang, X. (2016). Genome-wide assessment of differential translations with ribosome profiling data. *Nature Communications*, 7(1):1–11.

Xu, H., Xiao, T., Chen, C.-H., Li, W., Meyer, C. A., Wu, Q., Wu, D., Cong, L., Zhang, F., Liu, J. S., et al. (2015). Sequence determinants of improved CRISPR sgRNA design. *Genome Research*, 25(8):1147–1157.

Xu, Y., Poggio, M., Jin, H. Y., Shi, Z., Forester, C. M., Wang, Y., Stumpf, C. R., Xue, L., Devericks, E., So, L., et al. (2019). Translation control of the immune checkpoint in cancer and its therapeutic targeting. *Nature Medicine*, 25(2):301–311.

Xu, Y. and Ruggero, D. (2020). The role of translation control in tumorigenesis and its therapeutic implications. *Annual Review of Cancer Biology*, 4:437–457.

Yang, H.-S., Jansen, A. P., Komar, A. A., Zheng, X., Merrick, W. C., Costes, S., Lockett, S. J., Sonenberg, N., and Colburn, N. H. (2003). The transformation suppressor Pdcd4 is a novel eukaryotic translation initiation factor 4A binding protein that inhibits translation. *Molecular and Cellular Biology*, 23(1):26–37.

Yang, X., Zhong, W., and Cao, R. (2020). Phosphorylation of the mRNA cap-binding protein eIF4E and cancer. *Cellular Signalling*, page 109689.

Yang, Y., Fan, X., Mao, M., Song, X., Wu, P., Zhang, Y., Jin, Y., Yang, Y., Chen, L.-L., Wang, Y., et al. (2017). Extensive translation of circular RNAs driven by N 6-methyladenosine. *Cell Research*, 27(5):626–641.

Yang, Y. and Wang, Z. (2019). IRES-mediated cap-independent translation, a path leading to hidden proteome. *Journal of Molecular Cell Biology*, 11(10):911–919.

Yeomans, A., Thirdborough, S. M., Valle-Argos, B., Linley, A., Krysov, S., Hidalgo, M. S., Leonard, E., Ishfaq, M., Wagner, S. D., Willis, A. E., et al. (2016). Engagement of the b-cell receptor of chronic lymphocytic leukemia cells drives global and MYC-specific mRNA translation. *Blood*, 127(4):449–457.

Yewdell, J. W., Antón, L. C., and Bennink, J. R. (1996). Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules? *The Journal of Immunology*, 157(5):1823–1826.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterprofiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287.

Zhai, W. and Comai, L. (2000). Repression of RNA polymerase i transcription by the tumor suppressor p53. *Molecular and Cellular Biology*, 20(16):5930.

Zhang, H., Dou, S., He, F., Luo, J., Wei, L., and Lu, J. (2018a). Genome-wide maps of ribosomal occupancy provide insights into adaptive evolution and regulatory roles of uORFs during Drosophila development. *PLoS Biology*, 16(7):e2003903.

Zhang, H., Wang, Y., and Lu, J. (2019). Function and evolution of upstream ORFs in eukaryotes. *Trends in Biochemical Sciences*, 44(9):782–794.

Zhang, H., Wang, Y., Wu, X., Tang, X., Wu, C., and Lu, J. (2021). Determinants of genome-wide distribution and evolution of uORFs in eukaryotes. *Nature Communications*, 12(1):1–17.

Zhang, J., Medeiros, L. J., and Young, K. H. (2018b). Cancer immunotherapy in diffuse large b-cell lymphoma. *Frontiers in oncology*, 8:351.

Zhang, T., Li, N., Sun, C., Jin, Y., and Sheng, X. (2020). MYC and the unfolded protein response in cancer: synthetic lethal partners in crime? *EMBO Molecular Medicine*, 12(5):e11845.

Zhao, J.-J., Lin, J., Lwin, T., Yang, H., Guo, J., Kong, W., Dessureault, S., Moscinski, L. C., Rezania, D., Dalton, W. S., et al. (2010). microRNA expression profile and identification of miR-29 as a prognostic marker and pathogenetic factor by targeting CDK6 in mantle cell lymphoma. *Blood*, 115(13):2630–2639.

Zhou, P., Blain, A. E., Newman, A. M., Zaka, M., Chagaluka, G., Adlar, F. R., Offor, U. T., Broadbent, C., Chaytor, L., Whitehead, A., et al. (2019). Sporadic and endemic Burkitt lymphoma have frequent FOXO1 mutations but distinct hotspots in the AKT recognition motif. *Blood Advances*, 3(14):2118–2127.

Zuberek, J., Wyslouch-Cieszynska, A., Niedzwiecka, A., Dadlez, M., Stepinski, J., Augustyniak, W., Gingras, A.-C., Zhang, Z., Burley, S. K., Sonenberg, N., et al. (2003). Phosphorylation of eIF4E attenuates its interaction with mRNA 5 cap analogs by electrostatic repulsion: intein-mediated protein ligation strategy to obtain phosphorylated protein. *RNA*, 9(1):52–61.